High-throughput sequencing-based characterisation of fermented foods and their impacts on host gut microbiota

A thesis presented to the National University of Ireland for the degree of

Doctor of Philosophy

By

Aaron Walsh MSc

Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

APC Microbiome Ireland, University College Cork, Cork, Ireland

Department of Microbiology, University College Cork, Cork, Ireland

July 2018

Research supervisors: Dr Paul Cotter, Dr Marcus Claesson and Dr Fiona Crispie

Table of Contents

Chapter 2

Application of high-throughput sequencing technologies to study the microbiome of fermented foods and its relationship with flavour

Chapter 3

Microbial succession and flavour production in the fermented dairy beverage kefir

Chapter 4

Omics-based insights into flavour development and microbial succession within surface-ripened cheese

Chapter 5

Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks

Chapter 6

Species classifier choice is a key consideration when analysing low complexity food microbiome data

Chapter 7

A traditional fermented food modulates the murine gut microbiome while simultaneously ameliorating anxious-like behaviours in the animals

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD, is entirely my own work (except as declared, hereafter), and has not been submitted for another degree, either at University College Cork or elsewhere.

Signed:

Student Number: 109003070

Date: 09/07/2018

Thesis Abstract

Fermentation has been practised worldwide for millennia as a method to preserve or enhance foods, and, today, fermented foods remain a significant component in the human diet. Additionally, these foods are becoming increasingly popular since numerous health benefits have been ascribed to them, and thus it is necessary to (1) optimise their production, (2) assess their safety, and (3) determine the mechanisms by which they confer these effects. In this thesis, we examine if high-throughput sequencing technologies, particularly shotgun metagenomics, can address these needs. In Chapters 3 and 4, we show that shotgun metagenomics, when used alongside metabolomics, can be applied to understand the ways in which the microbiota influences flavour development in fermented foods. In Chapter 5, we report that shotgun metagenomics can accurately, and rapidly, detect pathogenic strains in fermented foods. In Chapter 6, we demonstrate that the choice of bioinformatics tools has a significant impact on shotgun metagenomic analysis of fermented foods. Finally, in Chapter 7, we provide evidence that a traditional fermented food modulates the gut microbiota in mice, while simultaneously reducing anxious-like behaviours in the animals. Overall, this thesis highlights that high-throughput sequencing is an invaluable tool for studying fermented foods. We illustrate that the technology not only expands our knowledge on the roles played by microorganisms during food fermentations, but it can also be used to ensure food safety or even investigate the ways in which these foods affect the host. Thus, high-throughput sequencing can bridge the gap between traditional food microbiology and health.

Publications

Walsh, Aaron M., Fiona Crispie, Kieran Kilcawley, Orla O'Sullivan, Maurice G. O'Sullivan, Marcus J. Claesson, and Paul D. Cotter. 2016. "Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir." *mSystems* no. 1 (5). doi: 10.1128/mSystems.00052-16.

Walsh, A. M., F. Crispie, M. J. Claesson, and P. D. Cotter. 2017. "Translating Omics to Food Microbiology." *Annu Rev Food Sci Technol* no. 8:113-134. doi: 10.1146/annurev-food-030216-025729.

Walsh, Aaron M., Fiona Crispie, Kareem Daari, Orla O'Sullivan, Jennifer C. Martin, Cornelius T. Arthur, Marcus J. Claesson, Karen P. Scott, and Paul D. Cotter. 2017. "Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks." *Applied and Environmental Microbiology*. doi: 10.1128/aem.01144-17.

Bertuzzi, A. S., A. M. Walsh, J. J. Sheehan, P. D. Cotter, F. Crispie, P. L. H. McSweeney, K. N. Kilcawley, and M. C. Rea. 2018. "Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese." *mSystems* no. 3 (1). doi: 10.1128/mSystems.00211-17.

Walsh, Aaron M., Fiona Crispie, Orla O'Sullivan, Laura Finnegan, Marcus J. Claesson, and Paul D. Cotter. 2018. "Species classifier choice is a key consideration when analysing low-complexity food microbiome data." *Microbiome* no. 6 (1):50. doi: 10.1186/s40168-018-0437-0.

List of Abbreviations

°C	Degrees Celcius
AMDIS	Automated Mass spectral Deconvolution and Identification System
amu	Atomic mass unit
ANOVA	Analysis of variance
APLSR	ANOVA-partial least squares regression
BH	Benjamini-Hochberg
bp	Base pair
cDNA	Complementary DNA
CLARK	CLAssifier based on Reduced K-mers
cm	Centimetre
CONCOCT	Clustering cONtigs on COverage and ComposiTion
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRW	Chinese rice wine
DIAMOND	Double index alignment of next-generation sequencing data
DNA	Deoxyribonucleic acid
dNTP	Nucleoside triphosphate
DVB/CAR/PDMS	Divinylbenzene/Carboxen/Polydimethylsiloxane
EC	Enzyme Commission
EMBA	Eosin methylene blue agar
ENA	European Nucleotide Archive
EPS	Exopolysaccharide
FAA	Free amino acid

FDA	Food and Drug Administration
FDR	False discovery rate
FFA	Free fatty acid
FMP	Fermented milk product
g	Gram
GABA	Gamma-aminobutyric acid
GB	Gigabyte
GC content	Guanine-cytosine content
GCMS	Gas chromatography–mass spectrometry
GI	Gastrointestinal
GO	Gene Ontology
GRAS	Generally regarded as safe
HGT	Horizontal gene transfer
HPLC	High performance liquid chromatorgraphy
HTS	High-throughput sequencing
HUMAnN2	HMP (Human Microbiome Project) Unified Metabolic Analysis Network
Hz	Hertz
IBD	Inflammatory bowel disease
IDBA-UD	Iterative De Bruijn graph de novo Assembler for short reads sequencing data with highly Uneven sequencing Depth
IMViC	Indole Methyl Red Vorges-Proskeur Citrate
ISO	International Organization for Standardization
ITS	Internal transcribed spacer
kb	Kilobase

LAB	Lactic acid bacteria
LDA	Linear disriminant analysis
LEfSe	LDA effect size
LRI	Linear retention index
m	Metre
MAP	Modified atmosphere packaging
MB	Megabyte
MDS	Multidimensional scaling
mg	Milligram
mgl	methionine gamma lyase
MG-RAST	The Metagenomics RAST (Rapid Annotation using Subsystem Technology) server
min	Minute
ml	Millilitre
MLST	Multilocus sequence typing
MLVA	Multiple locus variable number tandem repeat analysis
mm	Millimetre
mRNA	Messenger RNA
MRS	De Man, Rogosa and Sharpe
nmol	Nanomole
NSLAB	Non-starter lactic acid bacteria
OTU	Operational taxonomic unit
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCoA	Principal Coordinates Analysis

PCR	Polymerase chain reaction
PERMANOVA	Permutational analysis of variance
PFGE	Pulse field gel electrophoresis
PGM	Personal Genome Machine
PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
psi	Pound per square inch
QIIME	Quantitative Insights Into Microbial Ecology
qPCR	Quantitative PCR
QTLs	Quantitative trait loci
RDA	Ranking descriptive analysis
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
rpm	Revolutions per minute
rRNA	Ribosomal RNA
SBS	Sequencing by synthesis
ShET2	Shigella enterotoxin-2
SLIMM	Species Level Identification of Microorganisms from Metagenomes
SMRT	Single molecule real time [sequencing]
SNP	Single nucleotide polymorphism
SPME	Solid-phase microextraction
SRA	Short Read Archive
SSU	small subunit rRNA gene
ST	Sequence type
STEC	Shiga toxigenic *Escherichia coli*

SUPER-FOCUS	SUbsystems Profile by databasE Reduction using FOCUS (Find Organisms by Composition USage)
TB	Terabyte
TSI	Triple sugar iron
U/ml	Units per millilitre
VP	Vacuum packaging
w/v	weight/volume
WGS	Whole genome sequencing
WMS	Whole metagenome shotgun sequencing
xg	G-force
µl	Microlitre
µm	Micrometre
5HIAA	5-hydroxyindoleacetic acid
5-HT	Serotonin (5-hydroxytryptamine)

List of Figures

List of Tables

Chapter 1

Translating omics to food microbiology

Published in *Annual Review of Food Science and Technology*

(doi: https://doi.org/10.1146/annurev-food-030216-025729)

Authors: Aaron M. Walsh, Fiona Crispie, Marcus J. Claesson, and Paul D. Cotter

Contributions:

- **Candidate** wrote the review, with guidance from **FC**, **MJC**, and **PDC**

Abstract

This review examines the applications of omics technologies in food microbiology, with a primary focus on high-throughput sequencing (HTS) technologies. We discuss the different sequencing approaches applicable to the study of food-related microbial isolates and mixed microbial communities in foods, and we provide an overview of the sequencing platforms suitable for each approach. We highlight the potential for genomics, metagenomics, and metatranscriptomics to guide efforts to optimise food fermentations. Additionally, we explore the use of comparative and functional genomics to further our understanding of the mechanisms of probiotic action and we describe the applicability of HTS as a food safety measure. Finally, we consider the use of HTS to investigate the effects that ingested microbes have on the human gut microbiota.

INTRODUCTION

Over the past two decades, omics technologies have revolutionised biological research, and advances in DNA sequencing methods have been at the centre of this revolution (1). Since the first human genome sequence was published in 2001 (2), at an estimated cost of $3 billion, advances relating to high-throughput sequencing (HTS) platforms have resulted in an enormous decrease in sequencing costs and a corresponding increase in the number of published genomes (3). High-throughput sequencing has had a profound impact in microbiology, in particular, where it is used to determine the genome sequences of microbial isolates and overcome the limitations of culture-dependent analysis of microorganisms (4). In recent years, HTS has also yielded unprecedented insights into broader microbial populations within different environments (5-7), including many foods and food production facilities (8, 9). In this review, we describe how different HTS approaches are applied in food microbiology. Specifically, we explore how these methods can be used to study starter cultures and probiotics, understand the microbial dynamics of food fermentations and product spoilage, and to detect and trace outbreaks of foodborne pathogens. Through this process, the ways in which this knowledge has and will be used to improve the quality and safety of foods is highlighted.

OMICS APPROACHES APPLICABLE TO FOOD MICROBIOLOGY

Omics is an umbrella term that encompasses the HTS approaches (meta)genomics and (meta)transcriptomics, as well as metabolomics and (meta)proteomics (among others). Genomics can be defined as the generation and analysis of whole-genome sequences of DNA extracted from an organism (10). In comparative genomics,

bioinformatic analysis is used to evaluate differences between the whole-genome sequences of different organisms (11). In functional genomics, gene expression analysis or mutational analysis are used to predict the function of genes detected in an organism by whole-genome sequencing (12).

Metagenomics is a term that is often used to describe two different HTS approaches: amplicon sequencing and whole metagenome shotgun sequencing (WMS). In amplicon sequencing, marker-genes are PCR-amplified from DNA extracted from a mixed microbial community, sequenced and aligned against a reference database to determine the taxonomic composition of a sample. The most commonly used amplicon sequencing methods are 16S rRNA gene sequencing and ITS gene sequencing (hereafter referred to as 16S and ITS), which are used to profile bacterial and fungal communities, respectively (13, 14). Typically, amplicon sequencing is limited to genus-level identification, although some studies have achieved species-level assignments thanks to dedicated species classifiers and the use of longer read technologies (15-17). In contrast, in WMS, total genomic DNA extracted from a mixed microbial community is fragmented and sequenced to determine in a non-specific manner the entire (bacterial, eukaryotic and viral) gene content of a sample (18). WMS offers insights into the metabolic potential of a microbial community, and additionally, binning of metagenome sequences, using tools like CLARK, MetaPhlan2 and Kraken (19), can give species-level identification. Whole metagenome shotgun sequencing requires a higher sequencing depth than amplicon sequencing and, as a consequence, is more expensive (20).

Metatranscriptomics involves sequencing cDNA generated from mRNA transcripts extracted from a mixed microbial community to measure global gene expression in a

sample (18). Metatranscriptomics is technically challenging due to the unstable nature of mRNA and its underrepresentation relative to rRNA (21). In addition, it requires high-depth sequencing to detect differentially expressed transcripts present in low abundances (22). Consequently, metatranscriptomics is the most expensive of the HTS approaches (23).

In food microbiology, metabolomics and metaproteomics are employed for the identification and quantification of microbial metabolites and microbial proteins, respectively, within a food matrix (24, 25). In this review, we primarily focus on how HTS approaches can be used in food microbiology, but we do highlight instances where it is useful to integrate HTS approaches with metabolomics or metaproteomics.

OVERVIEW OF CURRENT SEQUENCING PLATFORMS

The first commercial HTS platform, the 454 Genome Sequencer, was released in 2005 (26). Since then, several other sequencers have been commercialised, including the HeliScope from Helicos and the SOLiD from ABI (27). At present, Illumina's range of sequencers (MiSeq, NextSeq 500, and the HiSeq series) and the Ion Torrent Personal Genome Machine (PGM) are the most commonly used sequencing platforms (28). The Illumina and Ion sequencers use different sequencing chemistries, but follow similar principles. Briefly, in Illumina sequencing, adaptor-ligated DNA fragments on the surface of a glass slide are amplified by bridge PCR to generate clusters. Subsequently, these clusters are sequenced using a sequencing-by-synthesis approach that involves cyclic rounds of single-base extension using a mixture of fluorescently labelled dNTPs and imaging to identify the incorporated

base (29). In contrast, in Ion sequencing, adaptor-ligated DNA fragments on the surface of beads are amplified by emulsion-PCR. Afterwards, the beads are deposited into micro-wells on a semiconductor sequencing chip, where a similar sequencing-by-synthesis reaction occurs. The incorporation of bases is detected by measuring pH changes caused by the release of hydrogen ions during DNA extension (30).

The Illumina and Ion sequencers each have their own pros and cons, and the choice of which sequencer to use in a study depends on the aims of the research (Figure 1) (27). The Illumina MiSeq and the Ion Torrent PGM are both suitable for amplicon sequencing, although the latter has a higher error rate (31). The Illumina NextSeq 500 and the Illumina HiSeq 2500 generate dramatically more data, 120 GB and 1 TB, respectively, than the Illumina MiSeq and the Ion Torrent PGM, 15 GB and 1 GB, respectively, and thus they are more suited to high-throughput applications, such as whole metagenome shotgun sequencing and metatranscriptomics (28). On the other hand, the Ion Torrent PGM has a considerably shorter run time than Illumina sequencers, which is useful in epidemiological investigations, for example (28, 32).

In addition, the PacBio RS II and the Oxford Nanopore MinION are important sequencing platforms. The PacBio uses single-molecule real-time (SMRT) sequencing technology (33), whereas the MinION uses nanopore sequencing technology (34). The PacBio and the MinION can generate 1 GB and 90 MB of data, with average sequencing read lengths of 14 kb and 6 kb, respectively (28). Importantly, the reads generated by the PacBio and the MinION are significantly longer than those generated by the Illumina and Ion sequencers (Table 1). Thus, the

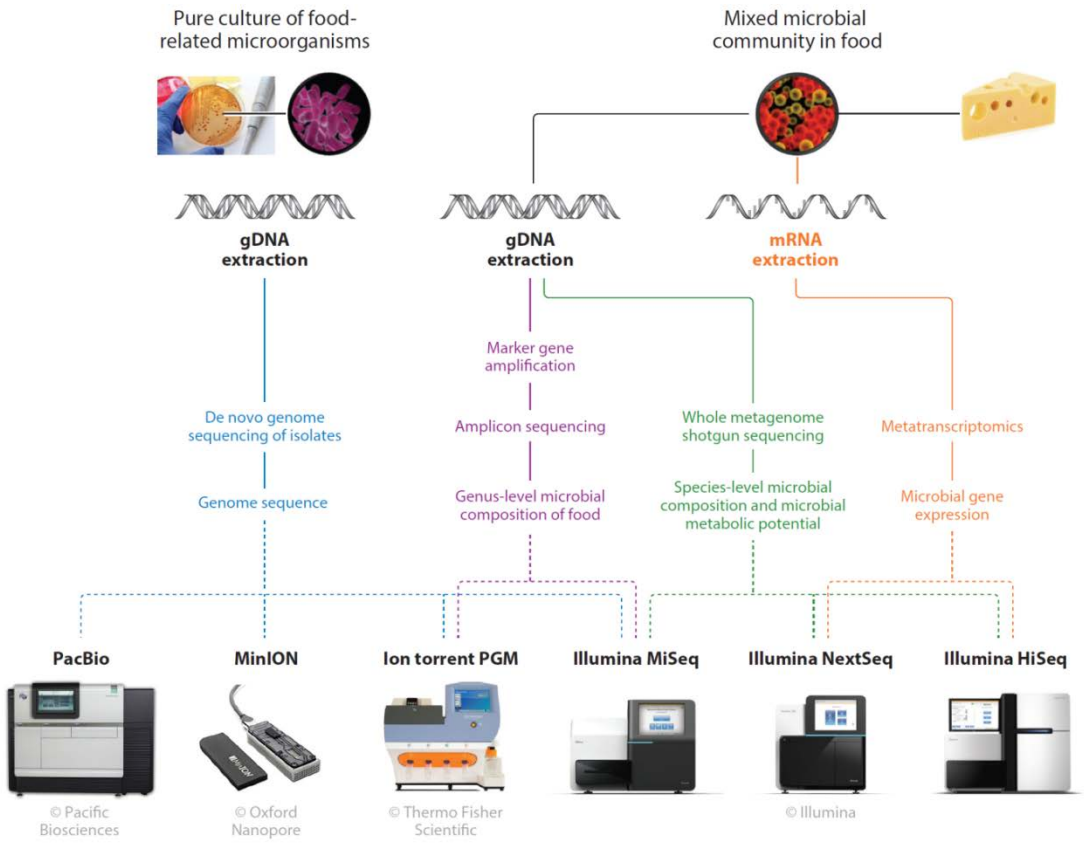


Figure 1: Schematic overview of the different high-throughput sequencing approaches applicable to food microbiology and suggestions for the sequencing platforms most suitable for each approach.

Table 1: Manufacturers' online specifications for the most commonly used sequencing platforms, as of June 2016.

Platform	Sequencing chemistry	Max output	Read length	Max no. of reads	Approximate run-time
Illumina MiSeq	Sequencing-by-synthesis (SBS)	15 GB	2 x 300 bp	25 million	4–55 hours
Illumina NextSeq	Sequencing-by-synthesis (SBS)	120 GB	2 × 150 bp	400 million	12–30 hours
Illumina HiSeq	Sequencing-by-synthesis (SBS)	1500 GB	2 × 150 bp	5 billion	1–3.5 days
Ion Torrent PGM	Semiconductor sequencing	2 GB	400 bp	5.5 million	2–7 hours
PacBio RS II	Single-molecule real-time (SMRT) sequencing	1 GB	14,000 bp	50 thousand	4 hours
MinION Mk I	Nanopore sequencing	90 MB	6,000 bp	16 thousand	18 hours

PacBio and MinION are ideal for *de novo* genome sequencing and complete assembly of microbial cultures (Figure 1) (35).

GENOMIC INSIGHTS INTO FOOD-RELATED MICROORGANISMS

Genomics of starter cultures

Fermentation has been practised worldwide for thousands of years to preserve foods and improve their nutritional and organoleptic properties (36). Since the end of the 19th century, preparations of one or more microorganisms called starter cultures have been used for the large-scale production of fermented foods with consistent qualities (37). Lactic acid bacteria (LAB), including *Lactobacillus, Lactococcus*, and *Streptococcus*, and yeasts, like *Saccharomyces cerevisiae*, are among the most important starter cultures (38), and to date, many different starters have had their genomes sequenced (39, 40).

Interestingly, the genome sequences of starter cultures have shed light on the history of their domestication (41). Comparative genomics has revealed that the adaptation of microorganisms to food coincided with gene-loss and gene-gain events (42). This was recently demonstrated by Zheng et al., who compared the genomes of gut and sourdough isolates of *Lactobacillus reuteri* to investigate how a microbe that was originally an intestinal symbiont adapted to a food environment (43). There was evidence for horizontal gene transfer and gene-loss events in sourdough isolates, and it was discovered that genes involved in energy metabolism and carbohydrate metabolism were more prevalent in these strains. The authors concluded that such genes might give sourdough isolates a competitive advantage during fermentation

(43). Such studies illustrate how an understanding of the evolution of starter cultures can highlight the genes that underpin the successes of microorganisms in fermented foods.

Traditionally, starter cultures were identified by screening natural isolates of microorganisms for desired traits (44). Bacteriophage immunity, exopolysaccharide (EPS) biosynthesis, and flavour formation are among the most important traits in starter cultures, and some of the genetic elements responsible for these traits are CRISPR/Cas loci, *eps* genes, and amino acid biosynthesis genes, respectively (45). Bioinformatics can be used to assist in the selection of starter cultures by screening the genomes of microorganisms for the presence of such genes (Betteridge, et al. 2015). Recently, the genomes of 213 industrial and natural *Lactobacillus* strains were sequenced (46). Bioinformatic analysis identified 48 glycoside hydrolase genes, important for sugar metabolism, and 60 cell envelope protein genes, important for flavour production, in the 213 genomes. Additionally, CRISPR/Cas loci were widespread in *Lactobacillus* isolates. Data generated by such large-scale sequencing projects can potentially be mined to identify candidate starter cultures.

Evolutionary engineering is another means of developing improved bacterial and fungal starter cultures (47). This technique involves continual propagation of microorganisms in vitro under a selective pressure to isolate mutants with desirable traits (48) and, for example, has been used to improve substrate utilisation and stress resistance in *Saccharomyces cerevisiae* (49). Similarly, 1,000 generations of continual propagation of a *Lactococcus lactis* strain in milk doubled its acidification rate (50). WGS can be used to identify the mutations responsible for the improved phenotypes achieved via evolutionary engineering (51).

Yeasts are central to the production of numerous foods, including bread, beer and wine (52). Unlike bacteria, yeasts are capable of sexual reproduction and can be crossed to breed new strains with enhanced characteristics (53). Many traits that are important in yeasts, like ethanol-tolerance or the production of aroma compounds, are under the control of multiple genes, known as quantitative trait loci (QTLs), which each contribute to the overall phenotype (54, 55). Genetic engineering is limited in its capacity to improve such complex traits and breeding is a more effective strategy (56). WGS can potentially be used to determine the presence of QTLs in yeasts to predict their breeding value (39). To date, this approach, known as genomic selection, has been under-utilised in yeasts, although it has shown considerable promise in cattle (57).

In addition to strain development, genome sequencing can be used to predict the metabolic requirements of starter cultures, and subsequently, this information can be used for fermentation optimisation (58). Indeed, this approach, known as metabolic modelling (59), has been used, for example, to develop minimal growth media for starters, identify alternative fermentable substrates, and improve amino acid production (60).

Genomics of probiotics

Probiotics are defined as microorganisms that confer health benefits when they are consumed in adequate amounts (61). Probiotics must reach the lower gastrointestinal (GI) tract alive to confer health benefits, and thus need to survive gastric transit and bile exposure. In addition, probiotics must adhere to the intestinal epithelia and mucosa to transiently colonise the gut (62). Several possible mechanisms have been

proposed to explain how probiotics confer health benefits, including the inhibition of pathogens via the production of antimicrobial proteins called bacteriocins or competitive exclusion of pathogens from the intestinal epithelia and mucosa, and by immunomodulation (63). Many probiotic strains have had their genomes sequenced (64) and this has given rise to the field of probiogenomics, wherein genomic approaches are used to understand how probiotics adapt to the gut and to explain how they exert health benefits (65).

Functional genomics has been used to determine the importance of individual genes to the mechanisms of probiotic action (62). For example, genome sequencing of the probiotic strain *Bifidobacterium breve* UCC2003 revealed that it had a gene-cluster encoding type IV Tad-pili, which had previously been shown to be involved in the adhesion of pathogens to the host (66). Disruption of the locus by insertional mutagenesis prevented mutants from colonising the murine gut, demonstrating that the pili are essential for host colonisation. Subsequent comparative genomics showed that the locus was conserved among *Bifidobacterium* genomes, suggesting that pili-mediated host colonisation is common to members of this genus (66). More recently, transposon mutagenesis was used to generate 1,110 *Lactobacillus casei* mutants, each with a mutation in a different gene, to identify those necessary for colonisation of the gut (67). In total, 47 genes that were essential for *L. casei* to colonise the ileal loop of rabbits were identified. These genes included some involved in housekeeping functions, cell wall synthesis, carbohydrate metabolism, amino acid metabolism, and environmental adaptation. Functional genomics has also been used to identify genes responsible for immunomodulation by probiotics. For instance, it was found that *L. casei* ATCC 27139 with a mutation in the gene *asnH* did not improve immunity in mice infected with *Listeria monocytogenes*, whereas the wild-type strain did (68).

The gene *asnH* encodes peptides that form part of the peptidoglycan layer, thereby suggesting that cell wall components of *L. casei* have a role in immunomodulation.

Going beyond the study of individual genomes, comparative genomics can explain why different probiotic strains have distinct effects. For example, comparative genomics was used to elucidate why some probiotic *Lactobacillus* species are associated with weight gain while others are associated with weight protection (69). Notably, the genomes of weight gain species did not encode enzymes necessary for fructose degradation, yet did encode enzymes which convert sucrose to fructose and glucose. In contrast, the genomes of weight protection species did encode enzymes necessary for fructose degradation and additionally encoded proteins involved in the synthesis of the anti-obesity compounds acetate, dextrin and L-rhamnose. Furthermore, the genomes of weight protection species contained glucose permease determinants. The authors suggest that the superior ability of weight protection species to degrade sugars reduces storage in the body, thus preventing weight gain. In addition, the genomes of weight gain species encoded thiolases, suggesting that these species may enhance fat digestion and fatty acid absorption/degradation, thus causing weight gain. Finally, the authors observed that weight protection species had more genes encoding bacteriocins than weight gain species (69). Comparative genomics has also been used to investigate the relationships between commercial probiotic strains. Comparison of the genomes of 34 *Lactobacillus acidophilus* strains, including multiple commercial strains, isolated over a 92 year period revealed that there was minimal genetic diversity in the species and strains shared almost identical genomes (70). This indicates that different *L. acidophilus* strains are likely to exert the same health benefits via the same mechanism.

As noted in the case of starter cultures above, genome sequencing has also been suggested as a method of identifying probiotic candidates, for example, by screening strains for genes encoding bile salt hydrolases or transporters, cell adhesins, and bacteriocins (71). Similarly, genome sequencing can be used to confirm the safety of probiotic candidates by screening for virulence genes or antibiotic resistance genes (72).

Meta-omics for the identification of probiotics

Many traditional fermented foods, like kefir and kimchi have been reported to have health benefits (73, 74), and an increasing number of metagenomes of the microbial populations of these foods have been sequenced, as discussed below. There is an opportunity to mine these metagenomes for strains with probiotic traits. For example, WMS analysis of kefir revealed that *Lactobacillus kefiranofaciens* had genes which encode proteins important for probiotic action, such as bile salt transporters, cell adhesins and bacteriocins (Walsh, et al. submitted).

Similarly, HTS analysis of the gut microbiota can be used to identify potential probiotics by highlighting correlations between the presence of particular microorganisms and the occurrence of diseases, like obesity and IBD, and such probiotics might be used in functional foods. The validity of this approach was recently demonstrated by Buffie et al., who administered antibiotics to mice to induce *Clostridium difficile* infection, and subsequently used 16S to examine the gut microbiota of mice that were resistant and susceptible to *C. difficile* (75). Correlation analysis indicated that 11 OTUs were associated with resistance in mice, including *Clostridium scindens*. Subsequent analysis of the gut microbiota of humans resistant

to *C. difficile* infection revealed that *C. scindens* was again associated with resistance. Thus, it was postulated that *C. scindens* protects the host against *C. difficile*. To confirm this, *C. scindens* was transferred to *C. difficile* infected mice, and the authors observed an amelioration of symptoms. Similar approaches can be adopted to discover probiotics to treat other diseases, as reviewed elsewhere (76).

Genomics of foodborne pathogens

Foodborne pathogens present a major public health concern. Annually, it is estimated that there are 9.4 million incidents of foodborne diseases in the United States alone, causing 56,961 hospitalisations and 1,351 deaths (77).

Whole genome sequencing (WGS) has revolutionised the field of epidemiology (78), and is particularly useful for investigating outbreaks of foodborne diseases. WGS allows epidemiologists to distinguish between outbreak and non-outbreak strains of foodborne pathogens by comparing the occurrence of single nucleotide polymorphisms (SNPs) in their genomes (79). It has been established that subtyping foodborne pathogens by WGS gives superior resolution to existing subtyping methods such as pulse-field gel electrophoresis (PFGE) and multiple-locus variable number tandem repeat analysis (MLVA) (80). Additionally, WGS of foodborne pathogens can be completed in a time-frame that is short enough for routine use in outbreak surveillance. In 2011, the Ion Torrent PGM was used to sequence the genome of the enterohemorrhagic *Escherichia coli* O104:H4 strain during an outbreak in Germany (32). It was reported that genome sequencing and assembly took just 62 hours. Since then, proof-of-concept studies have demonstrated that WGS can be used to detect outbreaks of Shiga toxin-producing *E. coli* O157 and

Salmonella enterica serovar Enteritidis (81, 82). Similarly, WGS was used to detect outbreaks of *Listeria monocytogenes* in a public health laboratory over a twelve-month period and the authors of the associated paper noted that this WGS-based approach was more effective than existing methods (83). This approach is being made ever more feasible by constantly improving technologies. For example, it has been demonstrated that a shorter, 6 hour, Illumina MiSeq run can be used to subtype *Salmonella* at the same resolution as a standard MiSeq run, and it was shown that a 2 hour Oxford Nanopore MinION run was sufficient to assign strains to an outbreak (84). In addition, bioinformatics tools have been developed to streamline the analysis of WGS data and allow faster subtyping of foodborne pathogens. The web-based tool SeqSero was recently developed to determine *Salmonella* serotypes from raw sequencing reads or assembled genomes (85). Furthermore, the FDA have established a database for the genomes of foodborne pathogens called GenomeTrakr with the aim of helping researchers to trace the food source of outbreaks (86).

Aside from disease surveillance, WGS can be employed to trace the sources and transmission routes of pathogens through the food-chain. In just one interesting example, WGS of *Escherichia coli* O157 isolates from cattle and sheep revealed that the same serotype infects both animals, suggesting that on-farm practices, like separating cattle and sheep, might help to prevent disease outbreaks (87).

Genomics of bacteriophage

Bacteriophage therapy has emerged as a novel strategy to prevent the contamination of foods by foodborne pathogens (88). Bioinformatic screening for virulence genes and antibiotic resistance genes in the genome sequences of viral candidates for

bacteriophage therapy can confirm that they are safe to use in foods (89). Indeed, analysis of the genome sequence of P100, a bacteriophage used to control *Listeria monocytogenes* in foods, helped regulators to grant it Generally Regarded as Safe (GRAS) status (90).

In contrast, in the dairy industry, bacteriophage infection of starter cultures is detrimental to food quality and often results in fermentation failure (91). *Siphoviridae* are the most common bacteriophage to infect dairy starters, and an increasing number of their genome sequences have been published (92). Bacteriophage genomics has provided novel insights into the mechanisms of interaction between bacteriophage and their hosts and, ultimately, this knowledge might enable the rational development of anti-phage measures (93).

META-OMICS INSIGHTS INTO MICROBIAL COMMUNITIES IN FOOD

In addition to sequencing the genomes of individual food-related microorganisms, HTS can be used to study mixed microbial communities in foods.

Amplicon sequencing

To date, the vast majority of HTS investigations of food microbiota have used amplicon sequencing, and a number of comprehensive reviews have summarised the findings of these studies (94-97). As discussed above, amplicon sequencing is used to determine the microbial composition of foods. Here, we will discuss a selection of studies that highlight how the insights yielded by amplicon sequencing can be practically applied to improve food quality.

Differences in the microbial communities of the same kinds of traditional fermented foods can cause significant variations in their organoleptic properties. 16S-based analysis of artisanal and commercial doenjang, a fermented soybean paste, identified variability in the bacterial populations of samples from different producers and revealed that the populations of commercial samples were simpler than those of artisanal samples (98). Such findings can inform the development of starter cultures for large-scale production of particular fermented foods with consistent qualities (97). In addition to starter culture selection, production practices greatly influence the microbial composition and flavour characteristics of fermented foods. 16S-based analysis of 62 Irish artisanal cheeses revealed that the bacterial composition of cheeses differed according to the type of milk and ingredients used in their production (99). More recently, 16S- and ITS-based analysis of the rinds of 137 cheeses from 10 different countries showed that the microbial composition varied between bloomy, natural and washed rind cheeses (100). It was discovered that geographic origin of the cheese did not affect the microbial composition of the rinds. Instead, it was observed that environmental conditions, especially moisture, correlated with differences in the microbial composition of the rinds (100). Such studies demonstrate that production practices can be manipulated to drive the formation of microbial communities to produce fermented foods with desired qualities. However, although it is possible to control fermentations to a certain extent, environmental microorganisms have an essential role in the production of fermented foods. 16S- and ITS-based analysis of cheese production facilities showed that *Debaryomyces* and *Lactococcus* species involved in fermentation dominated the surfaces of the facilities, indicating that microorganisms from the environment might contribute to the formation of the microbial communities in cheeses, thus affecting

their characteristics (101). Similarly, 16S- and ITS-based analysis demonstrated that the bacteria and yeasts present in sourdough bread were also prevalent on the surfaces of the bakery (102). Interestingly, 16S has shown that the bacteria on grapevines originate in soil, suggesting that differences in the soil microbiota of vineyards result in the distinct flavours of wines produced in different regions (103).

Food fermentation is a dynamic process involving continuous changes in microbial communities, and amplicon sequencing has been used to characterise these changes in numerous foods, including cheeses and meats (104, 105). Ultimately, such information can be used to identify biomarkers for the ripeness and quality of fermented foods. For example, Bokulich *et al.* were able to determine the standard of batches of American coolship ale based on 16S data (106).

As mentioned, fermentation can enhance the quality and shelf-life of foods. In contrast, food spoilage, caused by the production of undesirable microbial metabolites, is detrimental to the organoleptic properties of products (107). Several investigations have employed 16S to identify the bacteria responsible for spoilage in different foods. For example, 16S-based analysis of a range of spoiled foods, including meat, dairy, vegetable and egg products, revealed that psychrotrophic members of the genera *Lactobacillus, Lactococcus, Leuconostoc* and *Weissella* caused spoilage in all of those foods (108). Similarly, 16S revealed that psychrotrophic bacteria, probably originating from water reservoirs, were primarily responsible for spoilage in meats and seafood (109). The results of both studies indicate that refrigeration of foods selected for spoilage bacteria.

Spoilage can be minimised by improving hygiene practises in food production facilities. To date, several studies have used 16S to determine the sources of food

spoilage bacteria in food. For example, 16S was used to determine the origin of spoilage bacteria in beefsteaks (110). It was observed that all of the bacteria associated with spoiled beefsteaks were present in cattle carcasses, indicating that they were the main source of spoilers. Additionally, it was discovered that bacteria from the carcasses were able to establish on the surfaces of the abattoir, suggesting that the environment contributes to spoilage (110). Recently, 16S was used to characterise the bacterial composition of a sausage production facility and in sausage meat at different processing stages (111). Although spoilage-associated *Leuconostoc* species were present at low levels in raw meat, emulsion and plant surfaces, they were the most abundant species in spoiled sausages. The authors suggested that packaging and refrigeration selected for *Leuconostoc*. Interestingly, high levels of *Yersinia* species were detected on plant surfaces, although they accounted for less than 1% of the bacteria in spoiled sausages (111). In a similar study, 16S revealed that *Leuconostoc* species were the most abundant bacteria in ready-to-eat meals, and were a minor constituent of the microbiota of raw materials and processing plant surfaces (112). Finally, Bokulich *et al.* employed 16S and the Bayesian technique SourceTracker (113) to illustrate that raw materials were the most likely source of spoilage bacteria in a beer brewery (114). Thus, amplicon sequencing can be used to highlight stages in the production chain or areas in the production facility where improvements in hygiene practices are necessary.

In addition, 16S can be used to assess the impact of food-preservation measures on the growth of spoilage bacteria. For example, 16S was recently used to investigate the effect of nisin-packaging on the microbiota of beef burgers (115). The tool PICRUSt (116), which predicts the bacterial genetic content of a sample based on its compositional profile, indicated that nisin-packaging reduced the frequency of

metabolic pathways associated with spoilage, such as fatty acid biosynthesis pathways. Similarly, 16S was used to assess the effects that high oxygen modified-atmosphere packaging (MAP) and vacuum-packaging (VP) had on the microbiota of beef. It was observed that MAP spoiled ten days sooner than VP and 16S data revealed that there was a higher level of spoilage-associated *Leuconostoc* species in MAP than VP (117). Amplicon sequencing, therefore, can assist the selection of optimal food-preservation measures.

Beyond compositional analysis

The examples discussed above illustrate the usefulness of amplicon sequencing to study the microbiology of foods, and many published papers describe the use of this technique. However, there is a need to move beyond simply cataloguing the microorganisms that are present and instead focus on elucidating their roles (118). This can be achieved through whole metagenome shotgun sequencing (WMS; rather than the use of PICRUSt as a proxy), metatranscriptomics (RNA-Seq), and integrated omics approaches, which combine high-throughput sequencing with metabolomics or metaproteomics (18).

A number of studies have demonstrated that WMS can identify the microorganisms that are most important during fermentation. For example, WMS-based analysis of kimchi revealed that genes homologous with those in *Leuconostoc mesenteroides* subsp. *mesenteroides* ATCC 8293 and *Lactobacillus sakei* subsp. *sakei* 23K that are associated with the fermentation of carbohydrates, like mono- and oligosaccharides, were enriched in the kimchi microbiome, indicating an important role for those strains during kimchi fermentation (119). Interestingly, a high number of phage

DNA sequences were detected in kimchi, suggesting that bacteriophage infection might affect the microbial community dynamics during kimchi fermentation. Similarly, WMS-based analysis of a cocoa bean fermentation sample showed that genes associated with carbohydrate catabolism, especially heterolactic fermentation and pyruvate metabolism, were enriched in *Lactobacillaceae* (120). In addition, genes associated with pectinolysis and citrate metabolism were detected in *Enterobacteriaceae*, indicating that these bacteria might contribute to degradation of cocoa pulp and flavour formation, although this genus had not been considered to be important during cocoa pulp fermentation, previously (120). Furthermore, WMS can provide insights into the role of specific microorganisms in flavour production during fermentation. In the aforementioned 'cheese rind' study (100), WMS-based analysis of bloomy, natural, and washed cheese-rind microbial communities was also carried out and revealed that washed rind cheeses, noted for their pungent aromas, were enriched in a number of pathways involved in the production of flavour compounds. These included cysteine and methionine metabolism pathways, which are associated with the production of volatile sulphur compounds, and valine, leucine and isoleucine degradation pathways, which are associated with putrid and sweaty aromas (100). Furthermore, genes encoding lipases, proteases, and methionine-gamma-lyase, an important enzyme in the production of sulphur compounds which had only been found previously in *Brevibacterium linens*, were identified in *Pseudoalteromonas*, suggesting that this genus is involved in flavour production in cheese (100). Ultimately, such studies might guide the development of multi-strain starter cultures to produce foods with enhanced sensory qualities.

While WMS can be applied to guide approaches to enhance the qualities of fermented foods, it can also be employed to identify, and ultimately address,

microbes associated with defects. For example, WMS analysis of Chinese rice wine revealed that *Lactobacillus brevis* encodes genes that are potentially involved in spoilage, including genes associated with biotin synthesis, malolactic fermentation, and short-chain fatty acid production. Thus, the authors suggested that *L. brevis* might negatively impact the quality of CRW, and indeed, compositional analysis revealed that *L. brevis* was most prevalent in spoiled wine (121). Similarly, WMS was recently used to determine the causal agent of a pinking defect in cheeses (122). It was found that the thermophile, *Thermus thermophilus*, which had not previously been associated with the cheese microbiota, was enriched in defect cheeses and that associated genes involved in carotenoid production were enriched in these samples. Using this knowledge, the researchers proceeded to isolate *T. thermophilus* from defect cheeses. To confirm that this microbe caused pinking, the defect was recreated in a normal cheese by inoculating it with *T. thermophilus*, thus confirming that this species is responsible for the discoloration phenomenon. While this finding was important in its own right, it also highlights that this approach could be employed to identify the causes of other defects in cheeses, like flavour defects or late blowing (123), and eventually inform control-measures to prevent such defects.

WMS can be also be employed to detect pathogens in food, as demonstrated by Leonard et al., who employed this approach to detect *E. coli* in fresh spinach (124). However, it is not suitable for use in clinical laboratories because the error rates inherent to current sequencing platforms might lead to the misidentification of microbes, particularly at strain-level (125) but it is useful for investigating the transmission of pathogens through food production chains. In one instance, WMS was used to investigate how food processing affected the microbial composition of beef Although processing reduced the total number of bacteria in the meat, it was

noted that it resulted in an increase in the relative abundance of *Salmonella enterica*, *Escherichia coli* and *Clostridium botulinum*, potentially because of their ability to survive antimicrobial interventions (126). Thus, WMS can be used to identify the control points in the food production chain that best reduce contamination by foodborne pathogens.

WMS-based approaches are useful in many situations but are still limited in that they can only predict the metabolic capabilities of microorganisms. In contrast, RNA-Seq measures the extent to which different genes are transcribed, and thus it is a more informative method of elucidating their importance/roles in fermentations, as demonstrated in several recent studies. RNA-Seq-based analysis of kimchi revealed that genes associated with flavour production were expressed by *Leuconostoc mesenteroides* at the beginning of fermentation, suggesting that this species contributes to the organoleptic properties of kimchi (127). Similarly, a combined 16S/RNA-Seq-based analysis of the ripening of a traditional Italian Caciocavallo Silano cheese revealed strong correlations between the abundance of non-starter LAB (NSLAB) and the levels of expression of genes involved in amino acid and fatty acid catabolism, suggesting that these bacteria are important for cheese maturation (128). Likewise, RNA-Seq-based analysis of a surface-ripened cheese revealed that genes associated with proteolysis/lipolysis were highly expressed by *Geotrichum candidum,* indicating that this species is important for flavour production in the cheese (129). Furthermore, the authors identified a subset of genes that were differentially expressed across ripening stages, and they suggested that the ripeness of the cheese could be assessed by measuring the expression level of those genes. Similar studies have been conducted in Camembert and Reblochon cheeses (Lessard, et al. 2014, Monnet, et al. 2016).

To our knowledge, only one published paper has described combining HTS with metaproteomics to investigate the microbiology of a fermented food (130). In that study, 16S and ITS were used to characterise the bacterial and fungal composition of Pu-erh tea. In addition, liquid chromatography and mass spectrometry were used to identify the microbial proteins in the tea. It was observed that the bacterial community was dominated by Proteobacteria and the fungal community was dominated by the genus *Aspergillus.* 40 bacterial proteins and 295 fungal proteins were detected: 75% of the bacterial proteins were from Proteobacteria and 58.68% of the fungal proteins were from *Aspergillus.* 42 of the proteins detected were secreted or extracellular proteins, some of which (e.g. pectin lyase and cellobiohydrolase) could be involved in the degradation of tea leaves. These results provide direct evidence relating to the identity of the microorganisms, and associated proteins, that are involved in the fermentation of Pu-erh tea, and indicate that fungi are especially important in this process (130).

Another approach to combining multiple –omics technologies has involved combining HTS and metabolomics to reveal the microorganisms that are responsible for the production of metabolites, such as flavour compounds, in fermented foods. A number of investigations have used 16S and metabolomics to study microbial succession and the production of metabolites during the fermentation of several types of traditional Korean seafood. These studies identified correlations between *Halanaerobium* and acetate, butyrate and methylamines, thus suggesting that members of this genus are important for the production of those metabolites (131-133). Similarly, ITS and metabolomics were combined to demonstrate that changes in the fungal population of kombucha tea corresponded with increases in the levels of antioxidants over the course of 21-day fermentations, suggesting that fungi

contribute to the healthy characteristics of this beverage (134). Recently, an integrated amplicon sequencing and metabolomics approach was employed to investigate changes in the microbial composition and volatile profile of Zhenjiang aromatic vinegar during fermentation (135). The authors identified strong correlations between bacteria and volatiles but weaker correlations between fungi and volatiles, and thus they concluded that bacteria are more important for flavour production. Subsequent addition of *Acetobacter pasteurianus*, isolated from the vinegar, to the fermentation and caused increases in the levels of flavour compounds, including acetic acid and 2,3-butanediol, thus partially validating the correlations between bacteria and volatiles. (135). A major limitation of using this approach is that amplicon sequencing typically only gives genus-level resolution and so it is unable to detect species-level variations in the microbial composition of fermented foods, which can impact their quality (80). In contrast, WMS has been shown to give species-level resolution in a number of fermented foods, such as Mexican Cojita cheese and kefir grains (136, 137). We recently used WMS and metabolomics to identify associations between different species and flavour compounds in kefir milk (Walsh, et al. 2016, submitted), a traditional fermented beverage with reported health benefits (74). We identified strong positive correlations between *Acetobacter pasteurianus* and acetic acid, which is associated with vinegary flavours; *Lactobacillus kefiranofaciens* and carboxylic acids associated with cheesy flavours; *Leuconostoc mesenteroides* and diones associated with buttery flavours; and *Saccharomyces cerevisiae* and esters associated with fruity flavours. Our results suggest that integrating WMS and metabolomics has the potential to lead to the identification of the strains that merit inclusion as part of multi-strain starter cultures to produce traditional fermented foods with improved sensory qualities.

Integrated HTS and metabolomic approaches have also been used to enhance the safety of fermented foods by identifying the microorganisms responsible for the production of harmful metabolites. For example, 16S and metabolomics were used to identify the microbes responsible for the production of biogenic amines in Chinese rice wine (138). Strong correlations were found between the relative abundances of several genera and the levels of biogenic amines, but not between the *Lactobacillus* species present and biogenic amines. As a consequence, 30 *Lactobacillus* isolates were screened and *L. plantarum* JN01 was identified as one that did not produce biogenic amines and which was ethanol-tolerant and produced a low amount of acetic acid, indicating that it could be added to the rice wine without spoiling its flavour. Addition of this strain to the fermentation reduced the levels of biogenic amines in the wine by inhibiting the growth of other bacteria through the production of organic acids. Similar strategies can be adopted to increase the safety of other fermented foods.

META-OMIC INSIGHTS INTO THE EFFECTS OF INGESTED MICROBES ON THE GUT MICROBIOTA

An increasing number studies are using HTS to determine the effects that ingested microbes, either from fermented foods or probiotic products, have on the gut microbiota (Figure 2). The majority of these studies have employed 16S rRNA sequencing (139). For example, this approach was used to investigate the effects that three probiotic strains had on the gut microbiota of mice fed a high fat diet (140). It was found that administration of probiotics caused the gut microbiota to become more similar to that of mice fed a normal diet. Importantly, there was a reduction in

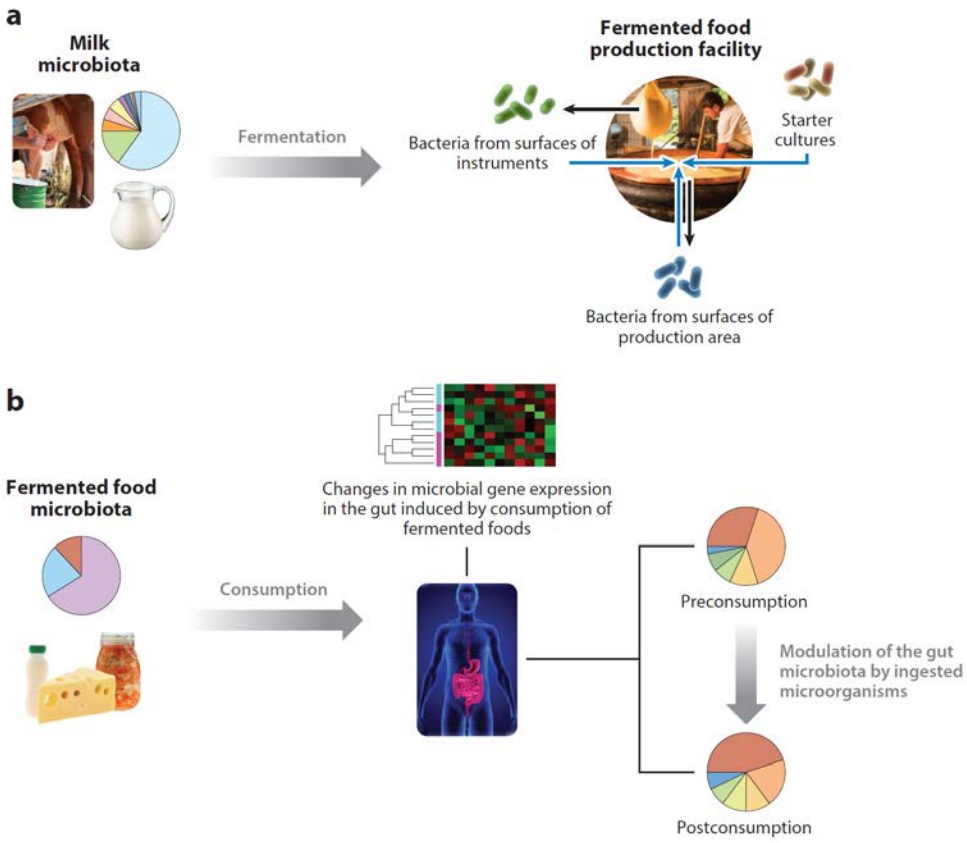


Figure 2: (a) Schematic overview of the flow of microorganisms through the food chain and (b) their impact on the human gut microbiota.

the levels of bacteria that were positively correlated with metabolic syndrome and an increase in the levels of bacteria that were negatively correlated with metabolic syndrome. In contrast, several studies have reported that ingestion of particular probiotics has no significant effects on the gut microbiota, based solely on 16S data (141-143). While this may be attributable to strain-specific effects, it is important to remember, as discussed above, that a major limitation of 16S is that it usually gives genus-level identification at best, and therefore it is not sensitive enough to detect potentially important changes at the species-level, or changes in gene expression. Thus, WMS or ideally RNA-Seq should be used for such investigations. McNulty et al. illustrated this when 16S rRNA showed that there was no change in the microbial composition of the gut microbiota of mice fed a fermented milk product (FMP), but RNA-Seq revealed that there was a significant increase in the expression of genes related to carbohydrate processing (144). Similarly, 16S rRNA sequencing showed that consumption of *Lactobacillus rhamnosus* LGG did not significantly alter the gut microbiota of elderly people, but RNA-Seq revealed there was an increase in the expression of anti-inflammatory pathways (145).

Although WMS is not as powerful as RNA-Seq, it has provided some invaluable insight into the reasons why fermented foods and probiotics may exert health benefits. For example, WMS was used to characterise changes at the species level in the microbiota of IBS patients fed an FMP (146). It was reported that FMP consumption induced an increase in the levels of anti-inflammatory butyrate-producing species, and a decrease in the levels of the pro-inflammatory species *Bilophila wadsworthia* and *Clostridium sp HGF_2*, which correlated with the alleviation of IBS symptoms (146). More recently, WMS was used to investigate how the probiotic mixture Prohep affects the microbial composition and gene

content of the gut microbiota of mice with hepatocellular carcinoma (147). It was observed that Prohep administration reduced tumour size by 40% and caused an increase in the abundance of bacteria and pathways that had anti-inflammatory effects, and a decrease in those with pro-inflammatory effects (147). In addition, HTS can be used to determine the fate of ingested microbes in the GI tract. This was recently demonstrated by David et al., who used 16S to show that there was a significant increase in the abundances of microbes originating from food in human subjects fed animal and plant based diets. Furthermore, RNA-Seq revealed that there was a significant increase in the gene expression of those microorganisms, indicating that they survive gastrointestinal transit and transiently colonise the gut (148). Finally, HTS can be used to predict the effect that ingested microbes will have on the gut microbiota, as shown by Zhang et al., who discovered that an FMP induced less changes in the gut microbiota of 'resistant' rats, which had a high abundance of indigenous *Lachnospiraceae*, than in 'permissive' rats, which had a low number of indigenous *Lachnospiraceae*, and they observed similar patterns in humans (149). They demonstrated that the 'resistant' and 'permissive' phenotypes could be replicated by faecal transplantation in gnotobiotic rats, confirming that the phenotypes were gut microbiota dependent. Their findings suggest that probiotic interventions should be tailored for individual patients, depending on their indigenous gut microbiota, and might explain why probiotic interventions are successful for some patients but not others.

OUTLOOK

High-throughput sequencing has transformed the field of food microbiology, enabling in-depth genomic characterisation of starter cultures, probiotics and foodborne pathogens, and additionally, culture-independent analysis of mixed microbial communities in foods and food production facilities. Indeed, whole genome sequencing of food-related microbial isolates has advanced to the point that it is routinely used to verify the safety of probiotic candidates and detect outbreaks of foodborne disease. While sequencing-based culture-independent analysis has also provided valuable insights, its use, and that of amplicon/compositional analysis in particular, is more limited. Typically, the short sequence reads generated by current sequencers provide limited resolution. However, strain-level variations between microorganisms can influence the organoleptic properties of foods, and thus strain-level resolution is more desirable. Although it has not been achieved to date, we expect that improvements in the throughput of long-read sequencers like the PacBio and MinION, in addition to Illumina synthetic long-read sequencing technology (150), or in combination with short read/high output sequencers, might allow strain-level resolution. Ultimately, we anticipate a time in the near future when it will be possible to use metagenomics to inform efforts to fine-tune fermentation processes and reliably test for the presence of foodborne pathogens.

References

1. **Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER.** 2013. The next-generation sequencing revolution and its impact on genomics. Cell **155:**27-38.
2. **Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W.** 2001. Initial sequencing and analysis of the human genome. Nature **409:**860-921.
3. **McPherson JD.** 2014. A defining decade in DNA sequencing. Nat Methods **11:**1003-1005.
4. **Loman NJ, Pallen MJ.** 2015. Twenty years of bacterial genome sequencing. Nature Reviews Microbiology.
5. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG.** 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences **109:**21390-21395.
6. **Consortium HMP.** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486:**207.
7. **Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF.** 2008. Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences **105:**3805-3810.
8. **Ercolini D.** 2013. High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology. Applied and Environmental Microbiology **79:**3148-3155.

9. **Bokulich NA, Lewis ZT, Boundy-Mills K, Mills DA.** 2016. A new perspective on microbial landscapes within food production. Current Opinion in Biotechnology **37:**182-189.

10. **Joyce AR, Palsson BØ.** 2006. The model organism as a system: integrating'omics' data sets. Nature Reviews Molecular Cell Biology **7:**198-210.

11. **Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S.** 2016. Ensembl comparative genomics resources. Database **2016:**bav096.

12. **Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M.** 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:**3674-3676.

13. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R.** 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences **108:**4516-4522.

14. **Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bolchacova E, Voigt K, Crous PW.** 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences **109:**6241-6246.

15. **Benitez-Paez A, Portune K, Sanz Y.** 2015. Species level resolution of 16S rRNA gene amplicons sequenced through MinIONTM portable nanopore sequencer. bioRxiv**:**021758.

16. **Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, Kaplan LA.** 2014. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. Journal of Microbiological Methods **104:**59-60.

17. **Allard G, Ryan FJ, Jeffery IB, Claesson MJ.** 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinformatics **16:**1.

18. **Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C.** 2015. Sequencing and beyond: integrating molecular'omics' for microbial community profiling. Nature Reviews Microbiology **13:**360-372.

19. **Lindgreen S, Adair KL, Gardner PP.** 2016. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports **6:**19233.

20. **Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R.** 2012. Unlocking the potential of metagenomics through replicated experimental design. Nature Biotechnology **30:**513-520.

21. **Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K, Meng J, Durham BP, Shen C, Varaljay VA.** 2013. Sizing up metatranscriptomics. The ISME Journal **7:**237-243.

22. **Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A.** 2011. Differential expression in RNA-seq: a matter of depth. Genome Research **21:**2213-2223.

23. **Lei R, Ye K, Gu Z, Sun X.** 2015. Diminishing returns in next-generation sequencing (NGS) transcriptome data. Gene **557:**82-87.

24. **Aldridge BB, Rhee KY.** 2014. Microbial metabolomics: innovation, application, insight. Current Opinion in Microbiology **19:**90-96.

25. **Hettich RL, Sharma R, Chourey K, Giannone RJ.** 2012. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Current Opinion in Microbiology **15:**373-380.

26. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-380.

27. **Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ.** 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology **30:**434-439.

28. **Reuter JA, Spacek DV, Snyder MP.** 2015. High-throughput sequencing technologies. Molecular Cell **58:**586-597.

29. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456:**53-59.

30. **Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M.** 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature **475:**348-352.

31. **Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG.** 2014. Performance comparison of Illumina and ion torrent next-generation

sequencing platforms for 16S rRNA-based bacterial community profiling. Applied and Environmental Microbiology **80:**7583-7591.

32. **Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W.** 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104: H4 outbreak by rapid next generation sequencing technology. PLoS One **6:**e22751.

33. **Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods **10:**563-569.

34. **Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J.** 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nature Biotechnology **33:**296-300.

35. **Koren S, Phillippy AM.** 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology **23:**110-120.

36. **Marsh AJ, Hill C, Ross RP, Cotter PD.** 2014. Fermented beverages with health-promoting potential: past and future perspectives. Trends in Food Science & Technology **38:**113-124.

37. **Caplice E, Fitzgerald GF.** 1999. Food fermentations: role of microorganisms in food production and preservation. International Journal of Food Microbiology **50:**131-149.

38. **Bourdichon F, Casaregola S, Farrokh C, Frisvad JC, Gerds ML, Hammes WP, Harnett J, Huys G, Laulund S, Ouwehand A.** 2012. Food fermentations: microorganisms with technological beneficial use. International Journal of Food Microbiology **154:**87-97.

39. **Borneman AR, Pretorius IS, Chambers PJ.** 2013. Comparative genomics: a revolutionary tool for wine yeast strain development. Current Opinion in Biotechnology **24:**192-199.

40. **Wu Q, Cheung CK, Shah NP.** 2015. Towards galactose accumulation in dairy foods fermented by conventional starter cultures: Challenges and strategies. Trends in Food Science & Technology **41:**24-36.

41. **Gibbons JG, Rinker DC.** 2015. The genomics of microbial domestication in the fermented food environment. Current Opinion in Genetics & Development **35:**1-8.

42. **Papadimitriou K, Pot B, Tsakalidou E.** 2015. How microbes adapt to a diversity of food niches. Current Opinion in Food Science **2:**29-35.

43. **Zheng J, Zhao X, Lin XB, Gänzle M.** 2015. Comparative genomics Lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. Scientific Reports **5:**18234.

44. **Betteridge A, Grbin P, Jiranek V.** 2015. Improving Oenococcus oeni to overcome challenges of wine malolactic fermentation. Trends in Biotechnology **33:**547-553.

45. **Kelleher P, Murphy J, Mahony J, van Sinderen D.** 2015. Next-generation sequencing as an approach to dairy starter selection. Dairy Science & Technology **95:**545-568.

46. **Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF.** 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nature Communications **6**.

47. **Çakar ZP, Seker UO, Tamerler C, Sonderegger M, Sauer U.** 2005. Evolutionary engineering of multiple-stress resistant Saccharomyces cerevisiae. FEMS Yeast Research **5:**569-578.

48. **Winkler JD, Kao KC.** 2014. Recent advances in the evolutionary engineering of industrial biocatalysts. Genomics **104:**406-411.

49. **Çakar ZP, Turanlı-Yıldız B, Alkım C, Yılmaz Ü.** 2012. Evolutionary engineering of Saccharomyces cerevisiae for improved industrially important properties. FEMS Yeast Research **12:**171-182.

50. **Bachmann H, Starrenburg MJ, Molenaar D, Kleerebezem M, van Hylckama Vlieg JE.** 2012. Microbial domestication signatures of Lactococcus lactis can be reproduced by experimental evolution. Genome Research **22:**115-124.

51. **Bachmann H, Pronk JT, Kleerebezem M, Teusink B.** 2015. Evolutionary engineering to enhance starter culture performance in food fermentations. Current Opinion in Biotechnology **32:**1-7.

52. **LEGRAS JL, Merdinoglu D, CORNUET J, Karst F.** 2007. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. Molecular Ecology **16:**2091-2102.

53. **Steensels J, Verstrepen KJ.** 2014. Taming wild yeast: potential of conventional and nonconventional yeasts in industrial fermentations. Annual Review of Microbiology **68:**61-80.

54. **Duitama J, Sánchez-Rodríguez A, Goovaerts A, Pulido-Tamayo S, Hubmann G, Foulquié-Moreno MR, Thevelein JM, Verstrepen KJ, Marchal K.** 2014. Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. BMC Genomics **15:**1.

55. **Steyer D, Ambroset C, Brion C, Claudel P, Delobel P, Sanchez I, Erny C, Blondin B, Karst F, Legras J-L.** 2012. QTL mapping of the production of wine aroma compounds by yeast. BMC Genomics **13:**573.

56. **Steensels J, Snoek T, Meersman E, Nicolino MP, Voordeckers K, Verstrepen KJ.** 2014. Improving industrial yeast strains: exploiting natural and artificial diversity. FEMS Microbiology Reviews **38:**947-995.

57. **Hayes BJ, Lewin HA, Goddard ME.** 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends in Genetics **29:**206-214.

58. **Smid E, Hugenholtz J.** 2010. Functional genomics for food fermentation processes. Annual Review of Food Science and Technology **1:**497-519.

59. **dos Santos FB, de Vos WM, Teusink B.** 2013. Towards metagenome-scale models for industrial applications—the case of Lactic Acid Bacteria. Current Opinion in Biotechnology **24:**200-206.

60. **Alkema W, Boekhorst J, Wels M, van Hijum SA.** 2015. Microbial bioinformatics for food safety and production. Briefings in Bioinformatics**:**bbv034.

61. **Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, Morelli L, Canani RB, Flint HJ, Salminen S.** 2014. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus

statement on the scope and appropriate use of the term probiotic. Nature Reviews Gastroenterology & Hepatology **11:**506-514.

62. **Bienenstock J, Gibson G, Klaenhammer TR, Walker WA, Neish AS.** 2013. New insights into probiotic mechanisms: a harvest from functional and metagenomic studies. Gut Microbes **4:**94-100.

63. **Bermudez-Brito M, Plaza-Díaz J, Muñoz-Quezada S, Gómez-Llorente C, Gil A.** 2012. Probiotic mechanisms of action. Annals of Nutrition and Metabolism **61:**160-174.

64. **Johnson BR, Klaenhammer TR.** 2014. Impact of genomics on the field of probiotic research: historical perspectives to modern paradigms. Antonie van Leeuwenhoek **106:**141-156.

65. **Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, O'Toole PW.** 2009. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nature Reviews Microbiology **7:**61-71.

66. **Motherway MOC, Zomer A, Leahy SC, Reunanen J, Bottacini F, Claesson MJ, O'Brien F, Flynn K, Casey PG, Munoz JAM.** 2011. Functional genome analysis of Bifidobacterium breve UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. Proceedings of the National Academy of Sciences **108:**11217-11222.

67. **Licandro-Seraut H, Scornec H, Pédron T, Cavin J-F, Sansonetti PJ.** 2014. Functional genomics of Lactobacillus casei establishment in the gut. Proceedings of the National Academy of Sciences **111:**E3101-E3109.

68. **Ito M, Kim Y-G, Tsuji H, Takahashi T, Kiwaki M, Nomoto K, Danbara H, Okada N.** 2014. Transposon Mutagenesis of Probiotic Lactobacillus casei

Identifies asnH, an Asparagine Synthetase Gene Involved in Its Immune-Activating Capacity. PLoS One **9:**e83876.

69. **Drissi F, Merhej V, Angelakis E, El Kaoutari A, Carrière F, Henrissat B, Raoult D.** 2014. Comparative genomics analysis of Lactobacillus species associated with weight gain or weight protection. Nutrition & Diabetes **4:**e109.

70. **Bull MJ, Jolley KA, Bray JE, Aerts M, Vandamme P, Maiden MC, Marchesi JR, Mahenthiralingam E.** 2014. The domestication of the probiotic bacterium Lactobacillus acidophilus. Scientific Reports **4**.

71. **Papadimitriou K, Zoumpopoulou G, Foligné B, Alexandraki V, Kazou M, Pot B, Tsakalidou E.** 2015. Discovering probiotic microorganisms: in vitro, in vivo, genetic and omics approaches. Frontiers in Microbiology **6**.

72. **Wei Y-X, Zhang Z-Y, Liu C, Malakar PK, Guo X-K.** 2012. Safety assessment of Bifidobacterium longum JDM301 based on complete genome sequences. World J Gastroenterol **18:**479-488.

73. **Park K-Y, Jeong J-K, Lee Y-E, Daily III JW.** 2014. Health benefits of kimchi (Korean fermented vegetables) as a probiotic food. Journal of Medicinal Food **17:**6-20.

74. **Bourrie BCT, Willing BP, Cotter PD.** 2016. The Microbiota and health promoting characteristics of the fermented beverage Kefir. Frontiers in microbiology **7**.

75. **Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L, Gobourne A, No D, Liu H, Kinnebrew M, Viale A.** 2015. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature **517:**205-208.

76. **Rebollar EA, Antwis RE, Becker MH, Belden LK, Bletz MC, Brucker RM, Harrison XA, Hughey MC, Kueneman JG, Loudon AH.** 2016. Using “omics” and integrated multi-omics approaches to guide probiotic selection to mitigate chytridiomycosis and other emerging infectious diseases. Frontiers in Microbiology **7:**68.

77. **Scallan E, Hoekstra R, Mahon B, Jones T, Griffin P.** 2015. An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. Epidemiology and Infection **143:**2795-2804.

78. **Kao RR, Haydon DT, Lycett SJ, Murcia PR.** 2014. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends in Microbiology **22:**282-291.

79. **Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S.** 2013. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med **10:**e1001387.

80. **Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, Boxrud D.** 2015. Characterization of foodborne outbreaks of Salmonella enterica serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. Journal of Clinical Microbiology **53:**3334-3340.

81. **den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z, Kinney A, Limberger R, Musser KA, Shudt M.** 2014. Rapid whole-

genome sequencing for surveillance of Salmonella enterica serovar enteritidis. Emerging Infectious Diseases **20:**1306.

82. **Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A.** 2015. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing Escherichia coli O157. Clinical Infectious Diseases**:**civ318.

83. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP.** 2015. Prospective whole genome sequencing enhances national surveillance of Listeria monocytogenes. Journal of clinical microbiology**:**JCM. 02344-02315.

84. **Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T.** 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol **16:**10.1186.

85. **Zhang S, Yin Y, Jones MB, Zhang Z, Kaiser BLD, Dinsmore BA, Fitzgerald C, Fields PI, Deng X.** 2015. Salmonella Serotype Determination Utilizing High-Throughput Genome Sequencing Data. Journal of clinical microbiology **53:**1685-1692.

86. **Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R.** 2016. the PRACTICAL value of Food Pathogen Traceability through BUILDING a Whole-Genome Sequencing Network and database. Journal of Clinical Microbiology**:**JCM. 00081-00016.

87. **Strachan NJ, Rotariu O, Lopes B, MacRae M, Fairley S, Laing C, Gannon V, Allison LJ, Hanson MF, Dallman T.** 2015. Whole Genome Sequencing demonstrates that Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association. Scientific reports **5**.

88. **Mahony J, van Sinderen D.** 2015. Novel strategies to prevent or exploit phages in fermentations, insights from phage–host interactions. Current Opinion in Biotechnology **32:**8-13.

89. **Endersen L, O'Mahony J, Hill C, Ross RP, McAuliffe O, Coffey A.** 2014. Phage therapy in the food industry. Annual Review of Food Science and Technology **5:**327-349.

90. **Chan BK, Abedon ST, Loc-Carrillo C.** 2013. Phage cocktails and the future of phage therapy. Future Microbiology **8:**769-783.

91. **Samson JE, Moineau S.** 2013. Bacteriophages in food fermentations: new frontiers in a continuous arms race. Annual Review of Food Science and Technology **4:**347-368.

92. **Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D.** 2016. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. Scientific Reports **6**.

93. **Mahony J, McDonnell B, Casey E, van Sinderen D.** 2016. Phage-Host Interactions of Cheese-Making Lactic Acid Bacteria. Annual Review of Food Science and Technology.

94. **Kergourlay G, Taminiau B, Daube G, Champomier Verges MC.** 2015. Metagenomic insights into the dynamics of microbial communities in food. Int J Food Microbiol **213:**31-39.

95. **Mayo B, Rachid CTC, Alegría Á, Leite AM, Peixoto RS, Delgado S.** 2014. Impact of next generation sequencing techniques in food microbiology. Current genomics **15:**293.

96. **Cocolin L, Ercolini D.** 2015. Zooming into food-associated microbial consortia: a 'cultural'evolution. Current Opinion in Food Science **2:**43-50.

97. **Galimberti A, Bruno A, Mezzasalma V, De Mattia F, Bruni I, Labra M.** 2015. Emerging DNA-based technologies to characterize food ecosystems. Food Research International **69:**424-433.

98. **Nam YD, Lee SY, Lim SI.** 2012. Microbial community analysis of Korean soybean pastes by next-generation sequencing. Int J Food Microbiol **155:**36-42.

99. **Quigley L, O'Sullivan O, Beresford TP, Ross RP, Fitzgerald GF, Cotter PD.** 2012. High-throughput sequencing for detection of subpopulations of bacteria not previously associated with artisanal cheeses. Applied and Environmental Microbiology **78:**5717-5723.

100. **Wolfe BE, Button JE, Santarelli M, Dutton RJ.** 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell **158:**422-433.

101. **Bokulich NA, Mills DA.** 2013. Facility-specific “house” microbiome drives microbial landscapes of artisan cheesemaking plants. Applied and Environmental Microbiology **79:**5214-5223.

102. **Minervini F, Lattanzi A, De Angelis M, Celano G, Gobbetti M.** 2015. House microbiotas as sources of lactic acid bacteria and yeasts in traditional Italian sourdoughs. Food Microbiology **52:**66-76.

103. **Zarraonaindia I, Owens SM, Weisenhorn P, West K, Hampton-Marcell J, Lax S, Bokulich NA, Mills DA, Martin G, Taghavi S.** 2015. The soil microbiome influences grapevine-associated microbiota. mBio **6:**e02527-02514.

104. **De Pasquale I, Di Cagno R, Buchin S, De Angelis M, Gobbetti M.** 2014. Microbial ecology dynamics reveal a succession in the core microbiota

involved in the ripening of pasta filata caciocavallo pugliese cheese. Applied and Environmental Microbiology **80:**6243-6255.

105. **Połka J, Rebecchi A, Pisacane V, Morelli L, Puglisi E.** 2015. Bacterial diversity in typical Italian salami at different ripening stages as revealed by high-throughput sequencing of 16S rRNA amplicons. Food Microbiology **46:**342-356.

106. **Bokulich NA, Bamforth CW, Mills DA.** 2012. Brewhouse-resident microbiota are responsible for multi-stage fermentation of American coolship ale. PLoS One **7:**e35507.

107. **Remenant B, Jaffrès E, Dousset X, Pilet M-F, Zagorec M.** 2015. Bacterial spoilers of food: Behavior, fitness and functional properties. Food Microbiology **45:**45-53.

108. **Pothakos V, Taminiau B, Huys G, Nezer C, Daube G, Devlieghere F.** 2014. Psychrotrophic lactic acid bacteria associated with production batch recalls and sporadic cases of early spoilage in Belgium between 2010 and 2014. International Journal of Food Microbiology **191:**157-163.

109. **Chaillou S, Chaulot-Talmon A, Caekebeke H, Cardinal M, Christieans S, Denis C, Desmonts MH, Dousset X, Feurer C, Hamon E.** 2014. Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. The ISME Journal.

110. **De Filippis F, La Storia A, Villani F, Ercolini D.** 2013. Exploring the sources of bacterial spoilers in beefsteaks by culture-independent high-throughput sequencing. PloS One **8:**e70222.

111. **Hultman J, Rahkila R, Ali J, Rousu J, Björkroth KJ.** 2015. Meat Processing Plant Microbiome and Contamination Patterns of Cold-Tolerant

Bacteria Causing Food Safety and Spoilage Risks in the Manufacture of Vacuum-Packaged Cooked Sausages. Applied and Environmental Microbiology **81:**7088-7097.

112. **Pothakos V, Stellato G, Ercolini D, Devlieghere F.** 2015. Processing Environment and Ingredients Are Both Sources of Leuconostoc gelidum, Which Emerges as a Major Spoiler in Ready-To-Eat Meals. Applied and Environmental Microbiology **81:**3529-3541.

113. **Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST.** 2011. Bayesian community-wide culture-independent microbial source tracking. Nature Methods **8:**761-763.

114. **Bokulich NA, Bergsveinson J, Ziola B, Mills DA.** 2015. Mapping microbial ecosystems and spoilage-gene flow in breweries highlights patterns of contamination and resistance. eLife **4:**e04634.

115. **Ferrocino I, Greppi A, La Storia A, Rantsiou K, Ercolini D, Cocolin L.** 2016. Impact of nisin-activated packaging on microbiota of beef burgers during storage. Applied and Environmental Microbiology **82:**549-559.

116. **Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R.** 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology **31:**814-821.

117. **Jääskeläinen E, Hultman J, Parshintsev J, Riekkola M-L, Björkroth J.** 2016. Development of spoilage bacterial community and volatile compounds in chilled beef under vacuum or high oxygen atmospheres. International Journal of Food Microbiology **223:**25-32.

118. **Hanage WP.** 2014. Microbiome science needs a healthy dose of scepticism. Nature **512:**247-248.

119. **Jung JY, Lee SH, Kim JM, Park MS, Bae J-W, Hahn Y, Madsen EL, Jeon CO.** 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. Applied and Environmental Microbiology **77:**2264-2274.

120. **Illeghems K, Weckx S, De Vuyst L.** 2015. Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. Food Microbiology **50:**54-63.

121. **Hong X, Chen J, Liu L, Wu H, Tan H, Xie G, Xu Q, Zou H, Yu W, Wang L.** 2016. Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. Scientific Reports **6:**26621.

122. **Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, Beresford TP, Ross RP, Fitzgerald GF, McSweeney PLH, Giblin L, Sheehan JJ, Cotter PD.** 2016. Thermus and the Pink Discoloration Defect in Cheese. mSystems **1**.

123. **O'Sullivan DJ, Giblin L, McSweeney PL, Sheehan JJ, Cotter PD.** 2013. Nucleic acid-based approaches to investigate microbial-related cheese quality defects. Frontiers in Microbiology 4 (2013): 1.

124. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2015. Application of metagenomic sequencing to food safety: detection of Shiga toxin-producing Escherichia coli on fresh bagged spinach. Applied and Environmental Microbiology **81:**8183-8191.

125. **Stasiewicz MJ, den Bakker HC, Wiedmann M.** 2015. Genomics tools in microbial food safety. Current Opinion in Food Science **4:**105-110.

126. **Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL.** 2016. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. Applied and Environmental Microbiology **82:**2433-2443.

127. **Jung JY, Lee SH, Jin HM, Hahn Y, Madsen EL, Jeon CO.** 2013. Metatranscriptomic analysis of lactic acid bacterial gene expression during kimchi fermentation. International Journal of Food Microbiology **163:**171-179.

128. **De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D.** 2016. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. Scientific Reports **6**.

129. **Dugat-Bony E, Straub C, Teissandier A, Onésime D, Loux V, Monnet C, Irlinger F, Landaud S, Leclercq-Perlat M-N, Bento P.** 2015. Overview of a surface-ripened cheese community functioning by meta-omics analyses. PloS One **10:**e0124360.

130. **Zhao M, Zhang D-l, Su X-q, Duan S-m, Wan J-q, Yuan W-x, Liu B-y, Ma Y, Pan Y-h.** 2015. An Integrated Metagenomics/Metaproteomics Investigation of the Microbial Communities and Enzymes in Solid-state Fermentation of Pu-erh tea. Scientific Reports **5**.

131. **Jeong SH, Jung JY, Lee SH, Jin HM, Jeon CO.** 2013. Microbial succession and metabolite changes during fermentation of dongchimi,

traditional Korean watery kimchi. International Journal of Food Microbiology **164:**46-53.

132. **Lee SH, Jung JY, Jeon CO.** 2014. Microbial successions and metabolite changes during fermentation of salted shrimp (saeu-jeot) with different salt concentrations. PLoS One **9:**e90115.

133. **Lee SH, Jung JY, Jeon CO.** 2015. Bacterial community dynamics and metabolite changes in myeolchi-aekjeot, a Korean traditional fermented fish sauce, during fermentation. International Journal of Food Microbiology **203:**15-22.

134. **Chakravorty S, Bhattacharya S, Chatzinotas A, Chakraborty W, Bhattacharya D, Gachhui R.** 2016. Kombucha tea fermentation: Microbial and biochemical dynamics. International Journal of Food Microbiology **220:**63-72.

135. **Wang Z-M, Lu Z-M, Shi J-S, Xu Z-H.** 2016. Exploring flavour-producing core microbiota in multispecies solid-state fermentation of traditional Chinese vinegar. Scientific Reports **6:**26818.

136. **Escobar-Zepeda A, Sanchez-Flores A, Baruch MQ.** 2016. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiology **57:**116-127.

137. **Nalbantoglu U, Cakar A, Dogan H, Abaci N, Ustek D, Sayood K, Can H.** 2014. Metagenomic analysis of the microbial community in kefir grains. Food Microbiology **41:**42-51.

138. **Liu SP, Yu JX, Wei XL, Ji ZW, Zhou ZL, Meng XY, Mao J.** 2016. Sequencing-based screening of functional microorganism to decrease the formation of biogenic amines in Chinese rice wine. Food Control **64:**98-104.

139. **Derrien M, van Hylckama Vlieg JE.** 2015. Fate, activity, and impact of ingested bacteria within the human gut microbiota. Trends in Microbiology.

140. **Wang J, Tang H, Zhang C, Zhao Y, Derrien M, Rocher E, Vlieg JEv-H, Strissel K, Zhao L, Obin M.** 2015. Modulation of gut microbiota during probiotic-mediated attenuation of metabolic syndrome in high fat diet-fed mice. The ISME Journal **9:**1-15.

141. **Kim S-W, Suda W, Kim S, Oshima K, Fukuda S, Ohno H, Morita H, Hattori M.** 2013. Robustness of gut microbiota of healthy adults in response to probiotic intervention revealed by high-throughput pyrosequencing. DNA Research **20:**241-253.

142. **Larsen N, Vogensen FK, Gøbel R, Michaelsen KF, Al-Soud WA, Sørensen SJ, Hansen LH, Jakobsen M.** 2011. Predominant genera of fecal microbiota in children with atopic dermatitis are not altered by intake of probiotic bacteria Lactobacillus acidophilus NCFM and Bifidobacterium animalis subsp. lactis Bi-07. FEMS Microbiology Ecology **75:**482-496.

143. **Roos S, Dicksved J, Tarasco V, Locatelli E, Ricceri F, Grandin U, Savino F.** 2013. 454 pyrosequencing analysis on faecal samples from a randomized DBPC trial of colicky infants treated with Lactobacillus reuteri DSM 17938. PLoS One **8:**e56710.

144. **McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, Goodman AL, Henrissat B, Oozeer R, Cools-Portier S, Gobert G.** 2011. The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. Science Translational Medicine **3:**106ra106-106ra106.

145. **Eloe-Fadrosh EA, Brady A, Crabtree J, Drabek EF, Ma B, Mahurkar A, Ravel J, Haverkamp M, Fiorino A-M, Botelho C.** 2015. Functional Dynamics of the Gut Microbiome in Elderly People during Probiotic Consumption. mBio **6:**e00231-00215.

146. **Veiga P, Pons N, Agrawal A, Oozeer R, Guyonnet D, Brazeilles R, Faurie J-M, van Hylckama Vlieg JE, Houghton LA, Whorwell PJ.** 2014. Changes of the human gut microbiome induced by a fermented milk product. Scientific Reports **4**.

147. **Li J, Sung CYJ, Lee N, Ni Y, Pihlajamäki J, Panagiotou G, El-Nezami H.** 2016. Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice. Proceedings of the National Academy of Sciences **113:**E1306-E1315.

148. **David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA.** 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature **505:**559-563.

149. **Zhang C, Derrien M, Levenez F, Brazeilles R, Ballal SA, Kim J, Degivry M-C, Quéré G, Garault P, van Hylckama Vlieg JE.** 2016. Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. The ISME Journal.

150. **Li R, Hsieh C-L, Young A, Zhang Z, Ren X, Zhao Z.** 2015. Illumina synthetic long read sequencing allows recovery of missing sequences even in the “finished” C. elegans genome. Scientific Reports **5**.

Chapter 2

Application of high-throughput sequencing technologies to study the microbiome of fermented foods and its relationship with flavour

Manuscript in preparation

Authors: Aaron M. Walsh, Christian Chervaux, Mickaël Boyer, and Paul D. Cotter.

Contributions:

- **Candidate** wrote the review with guidance from **CC**, **MB** and **PDC**

Abstract

The advent of high-throughput sequencing has enabled the study of the microbiota of fermented foods to an unprecedented degree. The technology allows the identification of microbes present within these foods, but it can also be used to predict or measure their activities during fermentation. Indeed, this ability to study microbial dynamics *in situ* has yielded invaluable insights into the ways in which microorganisms may contribute to qualities, especially flavour, in these foods. Here, current knowledge with respect to the fermented food microbiota, as gleaned from high-throughput sequencing-based analyses, is reviewed. Additionally, we highlight the many examples that demonstrate the potential for these technologies to reveal the ways in which microbes influence flavour development in these foods and, ultimately, guide efforts to modulate and improve food fermentations.

Introduction

Fermentation has been practiced for millennia as a means of food preservation or food quality enhancement (1). Today, fermented foods are also being increasingly consumed due to a greater appreciation of associated health benefits (2).

Food fermentation is the result of the biological activity of microbes present within food matrices (3). It is thus notable that the advent of high throughput DNA sequencing (HTS) has revolutionised food microbiology over the past decade by enabling high-quality culture-independent characterisation of microbial communities, including those present in fermented foods (4). A major motivation for such analyses is that an improved understanding on the microbiota within fermented foods might ultimately lead to enhanced food qualities, including sensory properties such as flavour.

Three different HTS approaches can be used to characterise the microbiota of fermented foods. These are: amplicon sequencing, whole metagenome shotgun sequencing, and metatranscriptomics (also known as RNA Seq). For amplicon sequencing, microbial DNA that has been extracted from a sample is PCR amplified using primers which facilitate the sequencing of hyper-variable regions within conserved marker genes. Next, the PCR products, or amplicons, are mapped against a marker gene database containing sequences representative of difference taxa. Such mapping ultimately allows one to estimate the proportions of the different taxa present within a sample. The most commonly used amplicon sequencing approach is 16S rRNA gene sequencing (5), which is used to profile the bacterial composition of

samples, while ITS gene sequencing (6) is commonly used to profile the fungal composition of samples.

Amplicon sequencing has been the most frequently used HTS approach for the characterisation of the microbiota of fermented foods (7, 8) (Figure 1). Although it has yielded many novel insights into the microbial diversity in these foods (9), amplicon sequencing has some inherent limitations. Firstly, it is typically limited to genus-level classification (10) and thus, importantly, it cannot account for variation in the microbiota at the species-level or strain-level. Secondly, it cannot provide a direct insight into the functions encoded by the microbes present in the sample. Therefore, amplicon sequencing offers limited insights into the roles played by different microbes in fermentations.

Shotgun metagenomics yields considerably more information than amplicon sequencing. For shotgun metagenomics, microbial DNA that has been extracted from a sample is randomly fragmented, and these DNA fragments are then sequenced. Shotgun metagenomic reads can be mapped against a functional database to reveal the genes or functions encoded by the microbiome, and they can also be mapped against a taxonomic database to profile the microbial composition of samples at high taxonomic resolutions, even at the strain-level (11). Shotgun metagenomics is more expensive than amplicon sequencing, since it necessitates a higher sequencing depth, in addition to greater computational costs (11). Consequently, shotgun metagenomics has been comparatively underutilised (7, 8) (Figure 1), but several recent studies have demonstrated the potential for this method to pinpoint ways to enhance food quality.

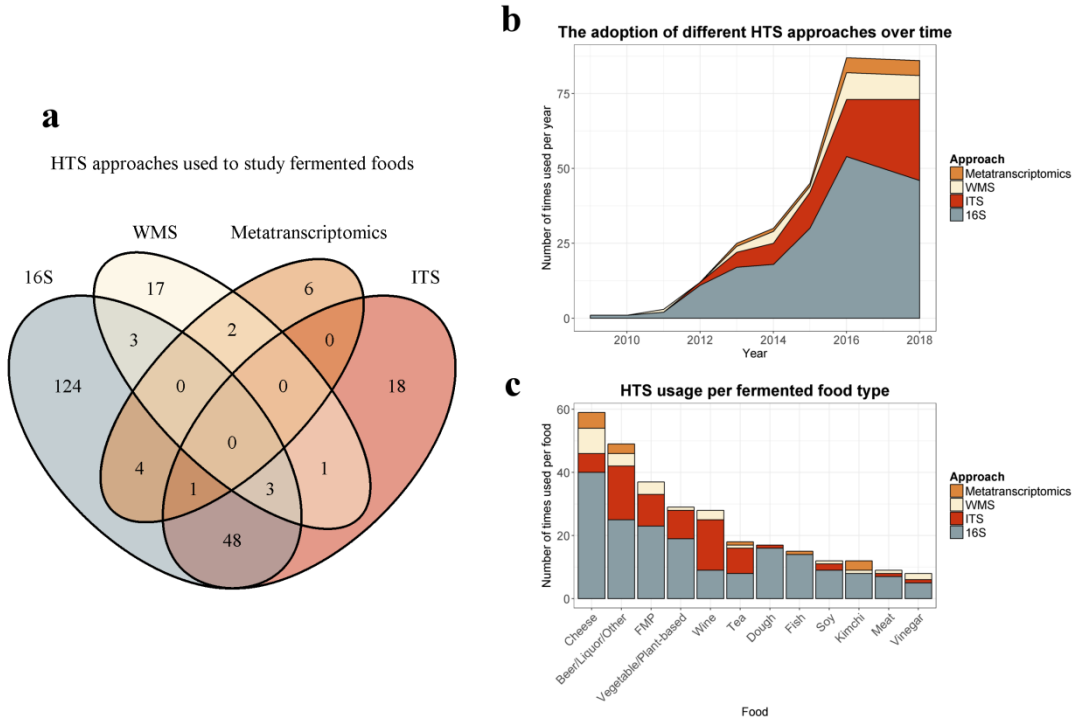


Figure 1: An overview on the usage of high-throughput sequencing (HTS) approaches for the analysis of fermented foods. (A) The Venn diagram shows the number of studies to adopt a given approach or combination thereof. (B) The stacked area chart shows the relative usage of approaches over time. (C) The stacked bar usage of HTS for different types of fermented foods. Note that this data was collected in January 2018.

Similarly, relatively few studies have used metatranscriptomics to study fermented food microbiota (7, 8) (Figure 1). For metatranscriptomics, cDNA synthesised from mRNA extracted from a sample is randomly fragmented, and these cDNA fragments are then sequenced. Metatranscriptomic reads are mapped against a functional database, and gene expression is measured by counting the number of reads which map to each gene. A considerable hurdle to metatranscriptomics is posed by the difficulty in isolating high quality mRNA from fermented foods, since mRNA is unstable, and thus it can degrade rapidly (12). Additionally, metatranscriptomics is significantly more expensive than either amplicon sequencing or shotgun metagenomics (13) as it requires very high sequencing depth. This issue is even more pronounced in instances where it is necessary to detect transcripts of genes that are expressed in low amounts. Regardless, metatranscriptomics is, potentially, an enormously useful sequencing approach for studying fermented foods, since it can determine levels of gene expression, and thus reflect the function of species, or even strains, during fermentation.

Each sequencing approach mentioned above can be used in conjunction with other omics methods, such as metabolomics or proteomics, to achieve multi-omics (14) analyses of fermented foods. Such analyses try to link changes in the proportions, functional potential or gene expression of microbes with biochemical changes that occur during food fermentations. Thus, multi-omics may help us to elucidate which microbes are important for particular organoleptic characteristics in fermented foods.

Here, we review studies that have used high-throughput sequencing, with an emphasis on shotgun metagenomics and metatranscriptomics, or multi-omic approaches to analyse common fermented foods. We discuss the ways in which the

information gained from such analyses might be applied to enhance food qualities like flavour. Additionally, we explore the potential for novel bioinformatics or computational biology methods to further our understanding of food fermentations.

Dairy

Fermented milk products (FMPs)

1. *Kefir*

Kefir is a traditional fermented milk beverage which originated in the Caucus region. It is produced by inoculating milk with a kefir grain. Afterwards, the milk is typically incubated at room temperature for around 24 hours. The kefir grains are cauliflower-like polysaccharide matrices harbouring a symbiotic community of bacteria and yeasts which are responsible for fermentation. Kefir is becoming increasingly popular due to reports of associated health benefits (15). HTS or multi-omic approaches can provide information that can be used to optimise the sensory properties of kefir, thus making it even more appealing to consumers.

Numerous studies have used amplicon sequencing to characterise the kefir microbiota. Early 16S rRNA gene sequencing studies revealed that kefir grains were dominated by *Lactobacillaceae* (16-18). Subsequent investigations have combined 16S rRNA gene sequencing with ITS gene sequencing to better characterise kefir. In one such instance, this approach was used to analyse 25 kefir grains, and their associated milks, which were sourced from 8 different countries (19). Within this study, 16S rRNA gene sequencing showed that the bacterial populations in the kefir

grains were dominated by *Lactobacillus*, but this genus was present at lower abundances in the milks, which were dominated by *Lactococcus*. Other subdominant bacterial genera detected included *Acetobacter* and *Leuconostoc.* Additionally, ITS gene sequencing established that the fungal populations in kefir were dominated by *Kazachstania*, *Kluyveromyces* and *Naumovozyma* (19). Other amplicon sequencing-based studies have since offered similar insights into the kefir microbiota, and they have consistently reported that *Lactobacillus kefiranofaciens* is the dominant bacterial species in kefir grains (20, 21).

Shotgun metagenomics and/or multi-omics have also been used to study kefir. Shotgun metagenomic analysis was first used to examine 2 kefir grains from Turkey (22). The authors found that the kefir grains were dominated by *Lactobacillus* species, specifically *Lactobacillus kefiranofaciens, Lactobacillus buchneri* and *Lactobacillus helveticus*. Additionally, they reported that the most abundant microbial pathways in the kefir grains were associated with the metabolism of carbohydrates, proteins, amino acids, and DNA or RNA (22). More recently, a multi-omics approach combining shotgun metagenomics with metabolomics was used to characterise changes in kefir milks from 3 countries over the course of 24 hour fermentations (23). Shotgun metagenomics revealed a consistent pattern of microbial succession across the 3 kefir milks. Specifically, *L. kefiranofaciens* was most abundant in the earlier stages but it decreased in later stages, when *Leuconostoc mesenteroides* increased. The observed changes in the microbial population corresponded with changes in the volatile profile of the kefir milks, and strong correlations were identified between the abundances of particular species and the levels of flavour compounds. Notably, *L. kefiranofaciens* correlated with carboxylic

acids and ketones, which are both associated with cheesy flavours, and esters, which are associated with fruity flavours. In contrast, *L. mesenteroides* correlated with 2,3-butanedione, which is associated with buttery flavours, and acetic acid, which is associated with vinegary flavours. The correlations indicated a causal relationship between the microbiota and the flavour of kefir, which was supported by evidence that spiking kefir with a *L. kefiranofaciens* isolate produced increases in ketones and esters, whereas spiking with a *L. mesenteroides* isolate produced increases in 2,3-butanedione and acetic acid. Additionally, sensory analysis showed that a kefir high in *L. mesenteroides* had a likeable buttery flavour, whereas another kefir high in *L. kefiranofaciens* had a less likeable but fruitier flavour. Interestingly, metagenome analysis also showed that *L. kefiranofaciens* lacked some pathways associated with aromatic amino acid biosynthesis, whereas *L. mesenteroides* contained these pathways. This is potentially important because there was a significant decrease in tyrosine during the fermentations. Thus, it is plausible that *L. mesenteroides* increased relative to *L. kefiranofaciens* due to its ability to synthesise tyrosine. The knowledge gained from this study might be used to manipulate the microbiota and, by extension, flavour of kefir, for example, by spiking kefir with isolates or modifying its nutrient content to favour the growth of particular microbes (23). A similar multi-omics approach, which instead combined amplicon sequencing with metabolomics, was used to study 4 kefir grains from different Turkish regions (24). Again, *L. kefiranofaciens* was linked to particular flavour compounds, like the ketone 2-butanone, which is associated with a yoghurt-like aroma. Furthermore, the kefir grains had similar but distinct microbial compositions, and the kefir milks had diverse aromatic profiles, and, thus, this provides further evidence that the kefir microbiota is important for its flavour.

2. Other traditional FMPs

Although kefir is perhaps the best studied traditional FMP, others have also been analysed using high-throughput sequencing methods. Several traditional FMPs produced by Mongolian peoples have been analysed using amplicon sequencing, including airag or koumiss (fermented mare's milk), khoormag (fermented camel's milk), and tarag (fermented cow's/goat's/yak's milk). Several studies using 16S rRNA gene sequencing have reported that *Lactobacillus* is the dominant bacteria across these milks (25-29). Additionally, ITS gene sequencing has indicated that *Galactomyces* is the dominant yeast in tarag (26), while *Pichia* was dominant in several naturally fermented cow milks produced by Mongolian people living in Russia (30).

Presently, only one published study describes the use of shotgun metagenomics to characterise a Mongolian FMP (29). Specifically, it was used to analyse 30 koumiss samples. The researchers reported that *Lactobacillus helveticus* was the dominant species, while *Lactococcus lactis, Lactobacillus buchneri, Lactobacillus kefiranofaciens* and *Acetobacter pasteurianus* were also prevalent species. The authors identified genes within the koumiss microbiome that are potentially important for flavour, including those associated with proteolysis. Additionally, a gene putatively encoding an aminotransferase, an enzyme involved in the transaminase pathway, was detected in koumiss. The transaminase pathway initiates the formation of many key flavour compounds, including aldehydes, organic acids, alcohols and esters. Furthermore sequences associated with amino lyases, which are involved in the production of sulphur compounds, were also observed in koumiss.

FMPs from other countries have also been analysed by high-throughput sequencing. The Colombian soured cream Suero Costeño was found to be dominated by either *Lactobacillus* or *Streptococcus* (31). A fermented goat's milk from China, called yond bap, was analysed by 16S rRNA gene sequencing, which revealed that most samples were dominated by *Pseudomonas* or *Lactococcus* (32). Furthermore, 16S rRNA gene sequencing revealed that a naturally fermented yak's milk from Tibet was dominated by the bacterial genus *Lactobacillus*, while ITS gene sequencing revealed that it was dominated the fungal genus *Saccharomyces* (33).

Matsoni, a popular FMP from the Caucuses, was also analysed by a combined amplicon sequencing approach (34). Overall, the most prevalent bacterial genera were *Lactobacillus* and *Streptococcus*, while the most prevalent fungal genera were *Kluyveromyces* and *Saccharomyces*. The authors reported significant variation, especially in the fungi, between matsonis produced using different milks or those from different regions, which indicated that production practices had a considerable influence on the matsoni microbiota. Additionally, they suggested that the unique flavours associated with regional matsonis may be attributable to their distinct regional microbiota.

Recently, a spontaneously fermented camel's milk from Ethiopia was analysed with 16S rRNA gene sequencing (35). *Streptococcus* was dominant but numerous potentially pathogenic genera were also prevalent, including *Escherichia* and *Klebsiella* (35). Similarly, shotgun metagenomics revealed that nunu, a spontaneously fermented cow's milk from Ghana, contained several potential pathogens (36). Notably, strain-level analysis of these samples detected an enterotoxin-producing *Escherichia coli* strain that was closely related to *E. coli*

O139:H28 E24377A, which had previously linked to a waterborne outbreak in India. Additionally, strain-level analysis detected an antibiotic resistant *Klebsiella pneumoniae* strain that was closely related to *K. pneumoniae* KpQ3, which had previously been linked a nosocomial outbreak among burn unit patients. Moreover, several undesirable functions were detected in the nunu metagenome, including histidine decarboxylases, which may produce biogenic amines, in addition to putrescine biosynthesis pathways, which may produce foul flavours.

Cheese

Cheese is the most widely consumed, and best studied, fermented dairy food. Many studies have used high-throughput sequencing, especially amplicon sequencing, to characterise microbial dynamics during curd fermentation or ripening, or microbial spatial distribution in cheeses, as reviewed elsewhere (7, 37). Here, we will focus on studies that have used shotgun metagenomics, metatranscriptomics, or multi-omics to study to study the microbial communities in cheese.

A seminal 2014 study used high-throughput sequencing to analyse rinds from 137 cheeses produced in 10 different countries (38). Amplicon sequencing revealed that the microbiota varied between bloomy, natural, and washed cheeses. Interestingly, the authors discovered that environmental conditions, especially moisture, had a significant influence on the cheese rind microbiota. Subsequent shotgun metagenomic analysis revealed that several pathways involved in flavour were enriched in washed rind cheeses, which are known for their particularly pungent aromas. Specifically, cysteine and methionine metabolism, which is associated with the production of sulphur compounds, was enriched in washed rind cheeses.

Additionally, isoleucine, leucine and valine degradation, which is associated with putrid or sweaty odours, was also enriched in these cheeses. Intriguingly, genes encoding enzymes important for flavour were identified in *Pseudoalteromonas*, including lipases, proteases, and methionine-gamma-lyase (*mgl*). Notably, *mgl*, which is involved in producing sulphur compounds, had only been found previously in *Brevibacterium linens.* Thus, *Pseudoalteromonas* might play a role in flavour development in cheese.

More recently, shotgun metagenomics was used to analyse Cotija, a cheese from Mexico (39). Here, it was shown that the Cotija metagenome contained genes associated with the production of many flavour compounds. The authors identified several transaminase genes that may transform free amino acids to alpha-keto acids, in addition to decarboxylases which may degrade these alpha-keto acids to aldehydes. Interestingly, no tryptophan or tyrosine transaminases were detected, which is important since their products, such as skatole, are associated with unappealing aromas. Additionally, complete fatty acid catabolism pathways, which produce methyl-ketones, were detected. Furthermore, genes encoding enzymes that convert methyl-ketones into their secondary alcohols were identified in the Cojita metagenome. Numerous genes encoding alcohol/aldehyde dehydrogenases, which are involved in aldehyde, carboxylate and ketone formation, were also found. Finally, the authors identified genes encoding enzymes which may enable microbes to synthesise alcohols from xylene in the cheese (39). Similarly, pathways which are potentially important for flavour formation, including those involved in proteolysis and amino acid catabolism, were detected in a washed-curd, brine-salted cheese using shotgun metagenomics (40).

Shotgun metagenomics can also been used to pinpoint the microbes responsible for defects in cheese quality. Indeed, shotgun metagenomics revealed that *Thermus thermophilus*, a thermophile which is not typically associated with the cheese microbiota, was enriched in cheeses with pink discoloration defect (41). Additionally, carotenoid biosynthesis genes were also enriched in those cheeses. Subsequently, the authors demonstrated that the pinking defect could be induced by adding *T. thermophilus* isolated from defected cheeses to normal cheeses, and thus they verified that this microbe caused the discoloration. Such studies illustrate that shotgun metagenomics may be utilised to identify the microbes responsible for other defects in cheeses, including flavour defects or late blowing (42), to ultimately inform control-measures to prevent their occurrence.

Metatranscriptomics was first used to study gene expression in an industrial Camembert-type cheese over a 77 day ripening period (43). It was observed that protease or peptidase genes were most highly expressed within the initial 21 days. The authors noted that genes associated with producing sulphur compounds were more highly expressed by the yeast *Geotrichum candidum* than the mould *Penicillium camemberti*. Conversely, genes associated with lipolysis were more highly expressed by *P. camemberti* than *G. candidum*. Overall, these findings suggest that the two fungi may contribute to distinct flavour characteristics in this cheese. Similarly, metatranscriptomics was used to study gene expression in a surface-ripened cheese over a 31 day ripening period (44). It was observed that *Lactococcus lactis* and *Kluyveromyces lactis* were the most active species on day 1. Subsequently, *Debaryomyces hansenii* and *Geotrichum candidum* became the most active species within the initial 14 days, and they remained dominant throughout

cheese maturation. Finally, acid-sensitive bacteria were active during the latter stages of cheese ripening. Genes associated with proteolysis in addition to lipolysis were mostly expressed by *G. candidum*, which suggests that this species might be central to flavour in this cheese. Interestingly, the authors detected genes which were differentially expressed at different maturation stages, and they proposed that these genes might be used as biomarkers to assess cheese ripeness (44). Another similar study used metatranscriptomics to study a Reblochon-style cheese during a 35 day ripening period (45). Again, *G. candidum* was reported to be the most metabolically active species during cheese maturation. Few changes were observed in bacterial gene expression, whereas there were changes in fungal gene expression. Notably, it was found that amino acid catabolism expression, including transaminase gene expression, increased by day 35, and these transcripts were attributed to *G. candidum* and *D. hansenii*.

Recently, metatranscriptomics was used to study microbial gene expression during ripening in a traditional Italian Caciocavallo Silano cheese (46). It was observed that transcripts related to amino acid metabolism and lipid metabolism, which are important for ripening, were enriched in the core whereas those related to carbohydrate metabolism were elevated on the crust. Additionally, it was investigated if ripening conditions influenced gene expression in the cheese. The authors reported that 651 genes were differentially expressed in the cores of cheeses ripened under higher temperatures compared to those ripened under standard temperatures. Indeed, numerous genes, including peptidases and lipases, and functions, including amino acid catabolism, fatty acid biosynthesis, and fatty acid beta-oxidation, that are associated with flavour were increased in the cores of the

cheeses ripened under higher temperatures. It was also found that genes involved in acetoin and diacetyl production were enriched in the crusts of these cheeses. The elevated amino acid metabolism expression seen in the cheeses ripened under higher temperatures was primarily attributed to Firmicutes. Interestingly, correlation analysis indicated that non-starter lactic acid bacteria (NSLAB), especially *Lactobacillus casei*, contributed significantly to the observed increase in amino acid metabolism expression, which suggested that NSLAB were important during cheese ripening. Moreover, the observed changes in gene expression within cheeses ripened under higher temperature coincided with increases in flavour compounds, as revealed by metabolomics, in addition to greater lipolysis and proteolysis indices (46).

A multi-omics approach, that combined shotgun metagenomics with metabolomics, was recently used to study surface-ripened cheeses, which were produced by smearing cheddar curd with commercial starter mixes, during a 30 day ripening period (47). Here, the authors observed consistent patterns in microbial succession within these cheeses, wherein yeast species like *D. hansenii* and *G. candidum* dominated during the initial stages, whereas bacterial species like *B. linens* and *Glutamicibacter arilaitensis* were more prevalent during the latter stages. Surface-ripened cheeses are noted for their intense flavours, and it was found that several pathways which are associated with flavour development, including lipolytic and proteolytic pathways, were significantly higher in the smeared cheeses than in an unsmeared cheese ripened under vacuum. Additionally, several strong correlations were identified between the relative abundances of individual species and the levels of particular flavour compounds in the cheeses. Specifically, *D. hansenii* correlated

with alcohols and carboxylic acids; *G. arilaitensis* correlated with alcohols, carboxylic acids and ketones; while *B. linens* and *G. candidum* correlated with sulphur compounds. Importantly, these correlations were supported by evidence from prior studies which had shown that these species can produce such compounds. Interestingly, *Staphylococcus xylosus*, which had only previously been associated with sulphur compounds in meats, was also found to correlate with sulphur compounds in the surface-ripened cheeses.

Plant-based fermented foods

Many plant-based fermented foods have been analysed by HTS. Here, we will focus on kimchi and soy, since HTS analyses have provided particularly valuable insights into the potential contributions of microbes to flavour in these foods.

Kimchi

Kimchi is a traditional fermented vegetable food from Korea. It is usually produced from cabbage or radish, while other ingredients, including spices, are often added for seasoning. Kimchi has been linked to numerous health benefits (48), and it is becoming increasingly consumed worldwide.

16S rRNA gene sequencing studies have established that kimchi is typically dominated by the genera *Lactobacillus*, *Leuconostoc*, and *Weissella* (49, 50). Generally, pH-sensitive *Leuconostoc* species are the most prevalent bacteria during the initial stages of kimchi fermentation, whereas the more pH-tolerant *Lactobacillus* and *Weissella* species become dominant as acidity increases in the latter stages. A

recent large-scale analysis of 88 kimchi revealed that there was some variability in the kimchi microbiota, which was attributed to factors like acidity, ingredients, and salinity (51). Several studies have demonstrated that changes in the kimchi microbiota correspond to changes in its metabolite profile, which indicates that bacteria are important for flavour development in kimchi (52-55). This was supported by a recent multi-omics analysis thatrevealed that lactic acid bacteria in kimchi produced 2-hydroxyisocaproic acid, a compound which has been associated with several benefits (56).

Shotgun metagenomic analysis of kimchi has also been carried out, focusing on a 29-day-fermentation (57). Genes associated with carbohydrate fermentation, especially saccharide fermentation, were found to be enriched in the kimchi metagenome. Most reads mapped to either *Lactobacillus sakei* subsp. *sakei* 23K or *L. mesenteroides* subsp. *mesenteroides* ATCC 8293, which suggests that these species drive kimchi fermentation. Indeed, changes in these species corresponded to changes in fermentation products, including mannitol, which is associated with a refreshing taste. Interestingly, it was also found that many reads mapped to phage genomes, which indicated that phage may influence the kimchi microbiota (57). Subsequently, the same authors used metatranscriptomics to analyse a subset of these kimchi samples (58). Here, it was confirmed that genes associated with heterolactic fermentation are central to kimchi fermentation. It was also observed that *Leuconostoc* species were the only ones to express mannitol dehydrogenase genes during kimchi fermentation, which indicated that *Leuconostoc* species were responsible mannitol production in kimchi. Additionally, metatranscriptomics yielded evidence that bacteria in kimchi may produce vitamins. Specifically, genes

associated with folate biosynthesis were expressed by *L. sakei*, while genes associated with riboflavin biosynthesis were expressed by *L. mesenteroides* (58). More recently, this data was reanalysed but with a specific focus on *L. mesenteroides* gene expression (59). Here, it was found that genes involved in the production of the flavour compounds acetoin, diacetyl and 2,3-butanediol were highly expressed in *L. mesenteroides*, thus providing further evidence that this species is important for flavour development in kimchi.

Soybean

Fermented soybean products are essential constituents of the Southeast Asian diet (60). Numerous studies have used HTS to study the microbiota of these foods.

Meju is an ingredient used to produce several traditional fermented soybean products from Korea. It is typically prepared by steaming soybeans that are then crushed to be moulded into blocks, which are fermented for one to two months under ambient conditions (60). 16S rRNA gene sequencing revealed that meju was dominated by the bacterial genus *Bacillus* throughout fermentation, but lactic acid bacteria were also prevalent (61), especially in the interior regions. Additionally, ITS gene sequencing revealed that meju is dominated by the fungal genus *Mucor* during the initial stages of fermentation, but it was dominated by *Aspergillus* during the latter stages (62).

Doenjang is a soybean paste that is produced by adding brine to meju, after which the mixture is fermented for up to 60 days (60). Doenjang has been associated with numerous health benefits, including anti-carcinogenic, anti-inflammatory, anti-

obesity, and anti-oxidant activities (63). 16S rRNA gene sequencing has revealed that doenjang contains the bacterial genera *Bacillus*, *Enterococcus, Lactobacillus, Leuconostoc, Staphylococcus* and *Tetragenococcus*, but their relative abundances vary between producers (64). A similar bacterial profile was detected in kochujang, which is a traditional Korean soybean paste that is made using meju powder (65), red pepper, and rice or glutinous rice flour (66).

Interestingly, 16S rRNA gene sequencing has been used in parallel with sensory analysis to associate the bacteria in doenjang with its sensory properties (67). Correlation analysis revealed that *Luteimonas, Ochrobactrum, Proteus, Rhodobacteraceae*, and *Stenotrophomonas* were found in doenjang with fermented fish sauce-like characteristics. In contrast, *Carnobacterium, Enterococcus, Pediococcus, Tetragenococcus*, and *Weissella* were associated with sourness. Finally, *Enterobacter* and *Enterococcus* were present in doenjang with a soft mouth-feel and a matured flavour, respectively (67). A recent multi-omics study combining 16S rRNA gene sequencing with metabolomics revealed that *Tetragenococcus* correlated with organic acids in doenjang, which indicated that this genus was driving the fermentation (68). It was also observed that *Lactobacillus* correlated with the bioactive compound gamma-aminobutyric acid in doenjang. Additionally, another multi-omics study indicated that *Lactobacillus* species in doenjang were associated with increased antioxidant activity, in addition to reduced cancer cell proliferation *in vitro* (69). Thus, *Lactobacillus* species may exert some of the health benefits associated with doenjang.

HTS has been used to study Chinese fermented soybean products, including Douchi (70) and soybean pastes (71). Notably, shotgun metagenomics was used to study

Chinese soy sauce over a 6-month-fermentation (72). It was found that the bacterial genus *Weisella* dominated during at the beginning of fermentation, whereas the fungal genus *Candida* dominated at the end. Interestingly, the increase in yeast coincided with increased ethanol production in Chinese soy sauce, in addition to a rise in genes associated with branched-chain amino acid metabolism, which suggests that yeasts were important for flavour development in this food.

Fermented tea

Kombucha

Kombucha is a fermented tea that is produced by adding a cellulosic pessicle, which is a mat containing a symbiotic microbial community, to sweetened tea, where it floats above the liquid (73). A new mat is formed following successful fermentation. Numerous health benefits have been attributed to kombucha (74), and consequently it is becoming increasingly popular in Western countries.

Several studies have combined 16S rRNA gene sequencing with ITS gene sequencing to characterise the kombucha microbiota. Notably, the first such study analysed five black tea kombuchas, which were produced using mats from four countries, over 10-day-fermentations (75). It was established that the dominant bacterial genus in the kombucha was *Gluconacetobacter*. The authors also observed that lactic acid bacteria, including *Lactobacillus*, were subdominant in kombucha, and their abundances increased during the fermentations. Additionally, it was found that the dominant fungal genus in the kombucha was *Zygosaccharomyces*, although *Dekkera* and *Kazachstania* were also detected.

Recently, amplicon sequencing was used to analyse kombucha that was produced by industrial-scale fermentations using either black tea or green tea (76). Here, the authors observed that black tea kombucha was dominated by the acetic acid bacteria *Gluconacetobacter*, whereas green tea kombucha was dominated the lactic acid bacteria *Oenococcus*. Correspondingly, acetic acid levels were highest in black teas, whereas lactic acid levels were highest in green teas. The tea type did not affect the yeast population, which was dominated by *Dekkera* and *Hanseniaspora* (76). However, another study revealed notable differences in the mycobiota of kombuchas that were produced using sterile tea, non-sterile tea, or honey tea (77).

16S rRNA gene sequencing has revealed that temperature also affects the bacterial composition of kombucha (78). It was found that the bacterial diversity of kombucha fermented at 30°C was greater than that of kombucha fermented at 20°C. Higher temperatures promoted the growth of lactic acid bacteria, including *Lactobacillus*, *Lactococcus*, and *Streptococcus*. The dominant genus in both kombuchas was *Gluconacetobacter*, but oligotyping (79) showed that different species were present at either temperature. Specifically, *Gluconacetobacter xylinus* was dominant at 20°C, whereas *Gluconacetobacter saccharivorans* was dominant at 30°C. The authors also reported that gluconic and glucuronic acids were higher at 30°C, and both acids were significantly correlated with *G. saccharivorans* (78).

Chakravorty *et al.* demonstrated that shifts in the kombucha microbiota during a 21-day-fermentation corresponded to increases in metabolites which are linked to health benefits (80). ITS gene sequencing revealed that *Candida* was dominant in the initial stages, but *Lachancea* became dominant at day 7. Meanwhile, biochemical analysis showed that flavonoids and polyphenols progressively increased during kombucha

fermentation. Additionally, fermentation augmented the anti-oxidant and anti-glycation activities of kombucha. Thus, the authors provided evidence that the kombucha microbiota may contribute to its health-promoting properties. To date, neither shotgun metagenomics nor metatranscriptomics have been applied to the study of the microbiota of kombucha.

Post-fermented teas

Post-fermented teas are produced *via* the solid-state fermentation of tea leaves by their endogenous microbes, and, as with kombucha, various health benefits have been linked to these teas (81). A number of post-fermented teas have been analysed with HTS methods, such as Fu-brick tea (82) or Liupao tea (83). However, to date, Pu-erh tea, which is produced in Yunnan in China (84), is the best characterised post-fermented tea. Several studies have established that Pu-erh tea is dominated by the bacterial phyla Actinobacteria, Firmicutes and Proteobacteria, and the fungal phylum Ascomycota (85, 86). Zhao *et al.* combined amplicon sequencing with metaproteomics for in-depth characterisation of Pu-erh tea (87). It was observed that the bacterial community was dominated by the phylum Proteobacteria, while the fungal community was dominated by the genus *Aspergillus*. Metaproteomic analysis of the tea identified 40 bacterial proteins, 75% of which were from Proteobacteria, and 295 fungal proteins, 58.68% of which were from *Aspergillus*. Additionally, 42 of the detected proteins were extracellular or secreted proteins, including some which may be important for the degradation of the tea leaves, such as cellobiohydrolase or pectin lyase. Thus, the authors provided evidence which suggested that microbes, and especially fungi, are central to Pu-erh tea fermentation (87). More recently, a study combining amplicon sequencing with metabolomics

revealed that changes in the microbiota correlated with changes in metabolites in the Pu-erh tea, which further emphasised the importance of the microbiota in this fermentation (88).

Sourdough

Sourdough bread is made from a flour-water mixture which is fermented by lactic acid bacteria and yeasts. These microbes produce organic acids which cause the pleasantly sour taste associated with this bread (89).

To date, amplicon sequencing is the only HTS approach that has been applied to study sourdoughs. Numerous 16S rRNA gene sequencing studies have reported that *Lactobacillus sanfranciensis* is the dominant species associated with sourdoughs, but other bacteria, such as *Enterococcus, Lactococcus, Leuconostoc, Pediococcus* and *Weissella*, are often found in these breads (90-92). Intriguingly, although *L. sanfransciensis* was dominant in sourdoughs from different French bakeries, it was found that these breads had distinct physiochemical characteristics, which led the authors to hypothesis that strain-level variation in *L. sanfransciensis* has a considerable impact on the sourdough qualities (93).

16S rRNA gene sequencing has also been used to assess the effects of ingredients on the sourdough microbiota. It has been reported that lactic acid bacteria were present at low abundances in flours used to produce sourdoughs (94), but they quickly became dominant after 1 day of sourdough fermentation (95, 96). Interestingly, it was observed that sourdoughs produced with rye flour were dominated by *Weissella*, whereas those produced with wheat flour were dominated by *Lactobacillus* (95).

Additionally, it was demonstrated that sourdoughs produced with organically farmed flour had a higher bacterial diversity than those produced by conventionally farmed flour (97). Conversely, sourdoughs produced with additional ingredients, like fruits or honey, had lower alpha diversity than those produced using normal ingredients (98).

Other studies have aimed to investigate the impact of fermentation conditions on the sourdough microbiota. Notably, amplicon sequencing was used to characterise the microbiota of 4 artisan sourdough bakeries in Italy (99). Here, the authors sampled sourdoughs, in addition to air, dough mixers, flours, storage boxes, and walls. It was observed that the same microbes which dominated the sourdoughs also dominated in the bakery. Indeed, 9 of the 11 detected bacterial OTUs and all of the detected fungal OTUs from the sourdoughs were shared with bakery equipment (99). Another study which used 16S rRNA gene sequencing revealed that the temperatures within sourdough bakeries impacted the breads' microbiota. Specifically, it was observed that sourdoughs fermented under controlled temperatures had a highly stable microbiota, whereas those fermented under ambient temperatures had a seasonally fluctuating microbiota (100).

Fermented seafood and meats

Fermented seafood is a staple in the Southeast Asian diet (101), and 16S rRNA sequencing studies have provided useful insights into the microbiota present in these foods. The first such study reported that 7 types of Korean fermented seafood were dominated by the bacterial genera *Lactobacillus* and/or *Weissalla,* and the archaeal family *Halobacteriaceae* (102). Similarly, 16S rRNA gene sequencing has revealed

that *Lactobacillus* is also prevalent in other fermented seafood products, such as Chinese Yucha (103), Japanese fermented sushi (104-106), and Korean gajami-sikhae (107). Unsurprisingly, it has been shown that salted-fermented seafood products, such as the shrimp paste saeu-jeot or the anchovy paste myeolchi-aekjeot, are dominated by halophilic bacteria. Interestingly, several studies have observed that *Halanaerobium* in these foods corresponded with increases in spoilage metabolites, including methylamines, which indicated that this genus may be detrimental to flavour. Recently, metatranscriptomics was used to characterise gene expression during salted-shrimp sauce fermentation (108). It was found that the halophile *Tetragenococcus halophilus* was the most metabolically active species at the studied time-point. Notably, transcripts associated with amino acid metabolism, peptidases and alpha-amylase were all assigned to *T. halophilus*, which led the authors to suggest that this species contributed to flavour in this food.

Fermentation has been practiced as a measure to preserve meat products since approximately 1500 BCE (109). The most popular fermented meat products, which mostly originated from Southern Europe, include fermented sausages, like chorizo from Spain or salami from Italy (110). HTS has been utilised to study the microbiota associated with fermented sausages, and amplicon sequencing has consistently indicated that *Lactobacillus* and *Staphylococcus* are generally the dominant bacteria in these foods (111-116). Recently, a multi-omics approach, which used shotgun metagenomics together with metabolomics, was employed to study Italian Felino salami that was fermented with or without starter cultures (117). Taxonomic analysis revealed that salami produced by inoculating meat with starter cultures had lower microbial diversity than salami produced by spontaneous fermentation, while

functional analysis identified 340 genes that were differentially abundant between the salami. Notably, genes encoding putative aldehyde reductases, acetate kinases, and 2,3-butanediol dehydrogenases, which are enzymes potentially involved in producing acetic acid, acetates, and acetoin, respectively, were higher in inoculated salami. Importantly, the metabolome reflected these differences in the metagenome, and it was observed that acetic acid, ethyl acetate and acetoin were indeed higher in inoculated salami. In contrast, genes associated with fatty acid biosynthesis were highest in spontaneously fermented salami, and, again, the metabolome reflected differences in the metagenome. Specifically, it was observed that long-chain esters, which can be derived from fatty acids, were higher in spontaneously fermented salami. Correlation analysis provided further evidence that the salami microbiota is pivotal to its flavour. Notably, it was demonstrated that different lactic acid bacteria correlated strongly with different esters: *Lactobacillus sakei* correlated with ethyl 2-methylbutanoate; *Lactococcus lactis* correlated with ethyl-alpha-hydroxybutyrate; *Lactobacillus brevis* correlated with ethyl esters; while *Leuconostoc citreum* correlated with ethyl isovalerate. Finally, sensory analysis revealed that inoculated salami was less likeable than spontaneously fermented salami, and the multi-omics data offered insights into the underlying reasons for this observation. The authors suggested that the genes which were enriched in the inoculated salami accelerated fermentation in those samples, and this caused acetic acid to be overproduced, and this, ultimately, negatively impacted flavour (117).

Alcoholic beverages

Alcoholic drinks, including beers, wines, and liquors, are the most widely consumed fermented beverages worldwide, and it has been understood since the 19th century

that microorganisms are essential in the production of these foods (118). Recently, integrated multi-omics approaches, utilising high-throughput sequencing in conjunction with metabolomics, have furthered our understanding of the role of the microbiota in flavour development in alcoholic beverages (119, 120).

Bokulich *et al.* used amplicon sequencing to characterise 200 commercial Californian wine fermentations (121). Interestingly, it was found that wine-producing areas, in addition to individual vineyards, could be distinguished by the microbial profile of their respective wines. Additionally, the regional differences in wine microbiota were closely correlated with differences in wine chemistry, which indicated that the microbiota had a significant influence on wine aroma. Indeed, a machine learning approach demonstrated that the microbiota could accurately predict the metabolome. Thus, the authors concluded that the microbiota may be used as a biomarker to assess wine quality (121).

Other studies have focussed on individual wines. Stefanini *et al.* used ITS gene sequencing to characterise the mycobiota in grape musts which are used to produce Amarone, a dry white wine from Italy, and it was found that the genus *Diplodia* positively correlated with flavour compounds (122). Elsewhere, a multi-omics approach was employed to analyse low-alcohol Merlot wines fermented with alternatives to the brewers' yeast *Saccharomyces cerevisiae* (123). Here, ITS gene sequencing indicated that these yeasts successfully propagated in the wine. Notably, wines fermented with *Metschnikowia pulcherrima* showed distinct metabolite profiles to those fermented with *S. cerevisiae*. Specifically, it was found that these wines contained high levels of esters along with sulphur compounds. Importantly, these wines also performed well in sensory evaluations.

Several studies have utilised HTS to link microorganisms to particular qualities in Chinese liquors. For example, shotgun metagenomics provided strong evidence that *Lactobacillus brevis* caused spoilage in a Chinese rice wine (124). Taxonomic analysis had revealed that this species was prevalent in spoiled rice wine, while functional analysis confirmed that it encoded genes, such as those associated with biotin biosynthesis or short-chain fatty acid production, which contribute to off-flavours (124). Another study, which utilised amplicon sequencing, discovered that the production facility environment was a major source of microbes during Chinese liquor fermentation (125). Interestingly, correlation analysis indicated that environmental microbes strongly influenced the metabolite profile of the liquor (125). Recently, Song *et al.* used a multi-omics approach to study Chinese Maotai-flavour liquor fermentation (126). Here, metabolomics revealed that the early stages were characterised by ethanol production, whereas the later stages were characterised by lactic acid production. Additionally, the microbial composition of the liquor was determined by amplicon sequencing. It was found that *Schizosaccharomyces* correlated with ethanol, while *Lactobacillus* correlated with lactic acid. Subsequently, metatranscriptomics confirmed the validity of these correlations. Briefly, it was demonstrated that *Schizosaccharomyces* expressed genes associated with ethanol production in the early stages. Specifically, the data indicated that *Schizosaccharomyces* converted pyruvate to acetaldehyde which it in turn converted to ethanol. It was also demonstrated that *Lactobacillus* expressed genes associated with lactic acid production in the later stages. Specifically, the data indicated that *Lactobacillus* converted pyruvate to lactic acid (126). Metatranscriptomics has also been used to measure the expression of genes associated with the production of two sulphur compounds, 3-(methylthio)-1-

propanol and dimethyl sulphide, which are important flavour compounds, in Chinese liquor (127). The authors reported that *Lactobacillus* and *Saccharomyces* were the most transcriptionally active microbes in the liquor. Importantly, it was observed that *Saccharomyces* was the only species to express every gene necessary to produce both compounds. However, it was noted that *Lactobacillus* expressed genes involved in recycling methionine, which is a precursor to both 3-(methylthio)-1-propanol and dimethyl sulphide. Thus, it was hypothesised that *Saccharomyces* and *Lactobacillus* may work synergistically to increase the production of these compounds. Subsequently, this was investigated *in vitro* by culturing *S. cerevisiae* and *L. buchnerii*, which were isolated from the liquor, in mono-culture or co-culture. It was found that *L. buchnerii* mono-cultures produced neither compound. Interestingly, though, it was demonstrated that co-cultures produced significantly more of the sulphur compounds than *S. cerevisiae* mono-cultures, thus confirming a synergistic relationship between these species.

Vinegar

Vinegar is a dilute solution of acetic acid which is used worldwide as a condiment or a pickling agent. It can be produced via the double fermentation of various sugary substrates, such as cereals or fruits, wherein ethanol is produced then subsequently converted to acetic acid (128). 16S rRNA gene sequencing has revealed that Chinese vinegars are dominated by *Lactobacillus* during the early stages, but *Acetobacter* increases over the course of fermentation (129, 130). The shift in the microbial community coincides with a decrease in ethanol, while there is a corresponding increase in acetic acid, and several studies have linked the vinegar microbiota to its flavour (131, 132). Notably, Wang *et al.* observed that *Acetobacter* correlated with

acetic acid, glutamic acid and 2,3-butanediol, and subsequent addition of *Acetobacter pasteurianus*, isolated from the vinegar, to the fermentation caused increases in these flavour compounds (132). More recently, shotgun metagenomics was used to identify the microbes responsible for acetoin production in Zhenjiang vinegar (133). The genetic pathway for diacetyl/acetoin production was reconstructed, and it was determined that *A. pasteurianus*, as well as four *Lactobacillus* species, potentially had the ability to synthesise acetoin from 2-acetolactate in the vinegar. The authors proceeded to isolate *A. pasteurianus* and three of the *Lactobacillus* species (*L. brevis*, *L. buchnerii*, and *L. fermentum*) from the vinegar. The isolates were then grown *in vitro* as co-cultures or mono-cultures. It was found that two co-cultures (*A. pasteurianus* plus *L. brevis* and *A. pasteurianus* plus *L. fermentum*) produced considerably more acetoin *in vitro* than mono-cultures did. Next, these two co-cultures were inoculated in vinegar, and it was observed that both caused a significant increase in acetoin *in situ* (133).

PART 3: Future directions and conclusions

This review highlights that integrated omics approaches, especially those utilising shotgun metagenomics or metatranscriptomics, have provided invaluable insights into the intricacies of microbial contributions to flavour development in fermented foods. Recently, exciting bioinformatics methods have been developed which have enormous potential to further extend our knowledge on these processes. Several tools have been released which enable strain-level analysis of microbiota (134) and, among these, PanPhlAn (135) and/or StrainEst (136) might be particularly useful to study fermented food microbiota. PanPhlAn aligns reads against a species-specific pangenome database to identify the gene families encoded by the strains in samples,

while StrainEst aligns reads against representative genomes to determine the single nucleotide variant profiles for the strains in samples. PanPhlAn can only detect the dominant strain from a species within samples, whereas StrainEst can detect multiple strains from the same species within samples. Both tools might be used to assess the effects of strain-level variation in fermented food microbiota on flavour development. Additionally, PanPhlAn, but not StrainEst, might also be used to examine changes in gene expression within strains over food fermentations to characterise their precise activities. Another potentially useful computational biology approach, aside from strain-level analysis, is metagenome-scale metabolic modelling, a method which uses the metagenome to predict which enzymes, and ultimately metabolites, may be produced by the microbiota (137). It has already been demonstrated that such an approach accurately predicted the metabolites produced by the gut microbiota in obese humans (138). Given the relative simplicity of fermented food microbiota, it is plausible that metagenome-scale metabolic modelling may be applied to these communities to predict the production of flavour compounds. Ultimately, the bioinformatics methods discussed here can improve our comprehension on the influence that strains exert on flavour development in fermented foods may guide starter culture optimisation. In conclusion, we expect that omics technologies enable an informed means to improve the flavour of fermented foods.

References

1. **Steinkraus KH.** 2002. Fermentations in World Food Processing. Comprehensive Reviews in Food Science and Food Safety **1:**23-32.
2. **Marco ML, Heeney D, Binda S, Cifelli CJ, Cotter PD, Foligné B, Gänzle M, Kort R, Pasin G, Pihlanto A, Smid EJ, Hutkins R.** 2017. Health benefits of fermented foods: microbiota and beyond. Current Opinion in Biotechnology **44:**94-102.
3. **Tamang JP, Watanabe K, Holzapfel WH.** 2016. Review: Diversity of Microorganisms in Global Fermented Foods and Beverages. Front Microbiol **7:**377.
4. **Walsh AM, Crispie F, Claesson MJ, Cotter PD.** 2017. Translating Omics to Food Microbiology. Annual Review of Food Science and Technology **8**.
5. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R.** 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences **108:**4516-4522.
6. **Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bolchacova E, Voigt K, Crous PW.** 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences **109:**6241-6246.
7. **De Filippis F, Parente E, Ercolini D.** 2017. Metagenomics insights into food fermentations. Microbial Biotechnology **10:**91-102.
8. **Cao Y, Fanning S, Proos S, Jordan K, Srikumar S.** 2017. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. Frontiers in Microbiology **8**.

9. **Kergourlay G, Taminiau B, Daube G, Champomier Verges MC.** 2015. Metagenomic insights into the dynamics of microbial communities in food. Int J Food Microbiol **213:**31-39.

10. **Noecker C, McNally CP, Eng A, Borenstein E.** 2017. High-resolution characterization of the human microbiome. Translational Research **179:**7-23.

11. **Quince C, Walker AW, Simpson JT, Loman NJ, Segata N.** 2017. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35:**833.

12. **Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K, Meng J, Durham BP, Shen C, Varaljay VA.** 2013. Sizing up metatranscriptomics. The ISME Journal **7:**237-243.

13. **Peñalver Bernabé B, Cralle L, Gilbert JA.** 2018. Systems biology of the human microbiome. Current Opinion in Biotechnology **51:**146-153.

14. **Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C.** 2015. Sequencing and beyond: integrating molecular'omics' for microbial community profiling. Nature Reviews Microbiology **13:**360-372.

15. **Bourrie BCT, Willing BP, Cotter PD.** 2016. The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir. Frontiers in Microbiology **7**.

16. **Dobson A, O'Sullivan O, Cotter PD, Ross P, Hill C.** 2011. High-throughput sequence-based analysis of the bacterial composition of kefir and an associated kefir grain. FEMS Microbiology Letters **320:**56-62.

17. **Leite A, Mayo B, Rachid C, Peixoto R, Silva J, Paschoalin V, Delgado S.** 2012. Assessment of the microbial diversity of Brazilian kefir grains by PCR-DGGE and pyrosequencing analysis. Food Microbiology **31:**215-221.

18. **Gao J, Gu F, He J, Xiao J, Chen Q, Ruan H, He G.** 2013. Metagenome analysis of bacterial diversity in Tibetan kefir grains. European Food Research and Technology **236:**549-556.

19. **Marsh AJ, O'Sullivan O, Hill C, Ross RP, Cotter PD.** 2013. Sequencing-based analysis of the bacterial and fungal composition of kefir grains and milks from multiple sources. PLoS One **8:**e69371.

20. **Korsak N, Taminiau B, Leclercq M, Nezer C, Crevecoeur S, Ferauche C, Detry E, Delcenserie V, Daube G.** 2015. Short communication: evaluation of the microbiota of kefir samples using metagenetic analysis targeting the 16S and 26S ribosomal DNA fragments. Journal of Dairy Science **98:**3684-3689.

21. **Garofalo C, Osimani A, Milanović V, Aquilanti L, De Filippis F, Stellato G, Di Mauro S, Turchetti B, Buzzini P, Ercolini D.** 2015. Bacteria and yeast microbiota in milk kefir grains from different Italian regions. Food Microbiology **49:**123-133.

22. **Nalbantoglu U, Cakar A, Dogan H, Abaci N, Ustek D, Sayood K, Can H.** 2014. Metagenomic analysis of the microbial community in kefir grains. Food Microbiology **41:**42-51.

23. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.

24. **Dertli E, Çon AH.** 2017. Microbial diversity of traditional kefir grains and their role on kefir aroma. LWT-Food Science and Technology **85:**151-157.

25. **OKI K, DUGERSUREN J, DEMBEREL S, WATANABE K.** 2014. Pyrosequencing analysis of the microbial diversity of airag, khoormog and tarag, traditional fermented dairy products of mongolia. Bioscience of Microbiota, Food and Health **33:**53-64.

26. **Sun Z, Liu W, Bao Q, Zhang J, Hou Q, Kwok L, Sun T, Zhang H.** 2014. Investigation of bacterial and fungal diversity in tarag using high-throughput sequencing. Journal of Dairy Science **97:**6085-6096.

27. **Gesudu Q, Zheng Y, Xi X, Hou QC, Xu H, Huang W, Zhang H, Menghe B, Liu W.** 2016. Investigating bacterial population structure and dynamics in traditional koumiss from Inner Mongolia using single molecule real-time sequencing. Journal of Dairy Science **99:**7852-7863.

28. **Zhong Z, Hou Q, Kwok L, Yu Z, Zheng Y, Sun Z, Menghe B, Zhang H.** 2016. Bacterial microbiota compositions of naturally fermented milk are shaped by both geographic origin and sample type. Journal of Dairy Science **99:**7832-7841.

29. **Yao G, Yu J, Hou Q, Hui W, Liu W, Kwok L-Y, Menghe B, Sun T, Zhang H, Zhang W.** 2017. A perspective study of koumiss microbiome by metagenomics analysis based on single-cell amplification technique. Frontiers in Microbiology **8**.

30. **Liu W, Zheng Y, Kwok L-Y, Sun Z, Zhang J, Guo Z, Hou Q, Menhe B, Zhang H.** 2015. High-throughput sequencing for the detection of the bacterial and fungal diversity in Mongolian naturally fermented cow's milk in Russia. BMC Microbiology **15:**45.

31. **Motato KE, Milani C, Ventura M, Valencia FE, Ruas-Madiedo P, Delgado S.** 2017. Bacterial diversity of the Colombian fermented milk

"Suero Costeño" assessed by culturing and high-throughput sequencing and DGGE analysis of 16S rRNA gene amplicons. Food Microbiology **68:**129-136.

32. **Liu X-F, Liu C-J, Zhang H-Y, Gong F-M, Luo Y-Y, Li X-R.** 2015. The bacterial community structure of yond bap, a traditional fermented goat milk product, from distinct Chinese regions. Dairy Science & Technology **95:**369-380.

33. **Liu W, Xi X, Sudu Q, Kwok L, Guo Z, Hou Q, Menhe B, Sun T, Zhang H.** 2015. High-throughput sequencing reveals microbial community diversity of Tibetan naturally fermented yak milk. Annals of Microbiology **65:**1741-1751.

34. **Bokulich NA, Amiranashvili L, Chitchyan K, Ghazanchyan N, Darbinyan K, Gagelidze N, Sadunishvili T, Goginyan V, Kvesitadze G, Torok T.** 2015. Microbial biogeography of the transnational fermented milk matsoni. Food Microbiology **50:**12-19.

35. **Fugl A, Berhe T, Kiran A, Hussain S, Laursen MF, Bahl MI, Hailu Y, Sørensen KI, Guya ME, Ipsen R.** 2017. Characterisation of lactic acid bacteria in spontaneously fermented camel milk and selection of strains for fermentation of camel milk. International Dairy Journal **73:**19-24.

36. **Walsh AM, Crispie F, Daari K, O'Sullivan O, Martin JC, Arthur CT, Claesson MJ, Scott KP, Cotter PD.** 2017. Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks. Applied and Environmental Microbiology**:**AEM. 01144-01117.

37. **Irlinger F, Layec S, Helinck S, Dugat-Bony E.** 2015. Cheese rind microbial communities: diversity, composition and origin. FEMS Microbiol Lett **362:**1-11.

38. **Wolfe BE, Button JE, Santarelli M, Dutton RJ.** 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell **158:**422-433.

39. **Escobar-Zepeda A, Sanchez-Flores A, Baruch MQ.** 2016. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiology **57:**116-127.

40. **Porcellato D, Skeie SB.** 2016. Bacterial dynamics and functional analysis of microbial metagenomes during ripening of Dutch-type cheese. International Dairy Journal **61:**182-188.

41. **Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, Beresford TP, Ross RP, Fitzgerald GF, McSweeney PL.** 2016. Thermus and the Pink Discoloration Defect in Cheese. mSystems **1:**e00023-00016.

42. **O'Sullivan DJ, Giblin L, McSweeney PL, Sheehan JJ, Cotter PD.** 2013. Nucleic acid-based approaches to investigate microbial-related cheese quality defects. Frontiers in Microbiology 4 (2013): 1.

43. **Lessard M-H, Viel C, Boyle B, St-Gelais D, Labrie S.** 2014. Metatranscriptome analysis of fungal strains Penicillium camemberti and Geotrichum candidum reveal cheese matrix breakdown and potential development of sensory properties of ripened Camembert-type cheese. BMC Genomics **15:**235.

44. **Dugat-Bony E, Straub C, Teissandier A, Onesime D, Loux V, Monnet C, Irlinger F, Landaud S, Leclercq-Perlat M-N, Bento P.** 2015. Overview of

a surface-ripened cheese community functioning by meta-omics analyses. PLoS One **10:**e0124360.

45. **Monnet C, Dugat-Bony E, Swennen D, Beckerich J-M, Irlinger F, Fraud S, Bonnarme P.** 2016. Investigation of the activity of the microorganisms in a Reblochon-style cheese by metatranscriptomic analysis. Frontiers in Microbiology **7:**536.

46. **De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D.** 2016. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. Scientific Reports **6**.

47. **Bertuzzi AS, Walsh AM, Sheehan JJ, Cotter PD, Crispie F, McSweeney PLH, Kilcawley KN, Rea MC.** 2018. Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese. mSystems **3**.

48. **Park K-Y, Jeong J-K, Lee Y-E, Daily III JW.** 2014. Health benefits of kimchi (Korean fermented vegetables) as a probiotic food. Journal of Medicinal Food **17:**6-20.

49. **Park E-J, Chun J, Cha C-J, Park W-S, Jeon CO, Bae J-W.** 2012. Bacterial community analysis during fermentation of ten representative kinds of kimchi with barcoded pyrosequencing. Food Microbiology **30:**197-204.

50. **Kyung KH, Medina Pradas E, Kim SG, Lee YJ, Kim KH, Choi JJ, Cho JH, Chung CH, Barrangou R, Breidt F.** 2015. Microbial ecology of watery kimchi. Journal of Food Science **80**.

51. **Lee M, Song JH, Jung MY, Lee SH, Chang JY.** 2017. Large-scale targeted metagenomics analysis of bacterial ecological changes in 88 kimchi samples during fermentation. Food Microbiology **66:**173-183.

52. **Jung JY, Lee SH, Lee HJ, Seo H-Y, Park W-S, Jeon CO.** 2012. Effects of Leuconostoc mesenteroides starter cultures on microbial communities and metabolites during kimchi fermentation. International Journal of Food Microbiology **153:**378-387.

53. **Jeong SH, Jung JY, Lee SH, Jin HM, Jeon CO.** 2013. Microbial succession and metabolite changes during fermentation of dongchimi, traditional Korean watery kimchi. International Journal of Food Microbiology **164:**46-53.

54. **Jeong SH, Lee HJ, Jung JY, Lee SH, Seo H-Y, Park W-S, Jeon CO.** 2013. Effects of red pepper powder on microbial communities and metabolites during kimchi fermentation. International Journal of Food Microbiology **160:**252-259.

55. **Jeong SH, Lee SH, Jung JY, Choi EJ, Jeon CO.** 2013. Microbial succession and metabolite changes during long-term storage of Kimchi. Journal of Food Science **78**.

56. **Park B, Hwang H, Chang JY, Hong SW, Lee SH, Jung MY, Sohn SO, Park HW, Lee JH.** 2017. Identification of 2-hydroxyisocaproic acid production in lactic acid bacteria and evaluation of microbial dynamics during kimchi ripening. Scientific Reports **7:**10904.

57. **Jung JY, Lee SH, Kim JM, Park MS, Bae J-W, Hahn Y, Madsen EL, Jeon CO.** 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. Applied and Environmental Microbiology **77:**2264-2274.

58. **Jung JY, Lee SH, Jin HM, Hahn Y, Madsen EL, Jeon CO.** 2013. Metatranscriptomic analysis of lactic acid bacterial gene expression during

kimchi fermentation. International Journal of Food Microbiology **163:**171-179.

59. **Chun BH, Kim KH, Jeon HH, Lee SH, Jeon CO.** 2017. Pan-genomic and transcriptomic analyses of Leuconostoc mesenteroides provide insights into its genomic and metabolic features and roles in kimchi fermentation. Scientific Reports **7:**11504.

60. **Shin D, Jeong D.** 2015. Korean traditional fermented soybean products: Jang. Journal of Ethnic Foods **2:**2-7.

61. **Kim Y-S, Kim M-C, Kwon S-W, Kim S-J, Park I-C, Ka J-O, Weon H-Y.** 2011. Analyses of bacterial communities in meju, a Korean traditional fermented soybean bricks, by cultivation-based and pyrosequencing methods. The Journal of Microbiology **49:**340-348.

62. **Jung JY, Lee SH, Jeon CO.** 2014. Microbial community dynamics during fermentation of doenjang-meju, traditional Korean fermented soybean. International Journal of Food Microbiology **185:**112-120.

63. **Patra JK, Das G, Paramithiotis S, Shin H-S.** 2016. Kimchi and Other Widely Consumed Traditional Fermented Foods of Korea: A Review. Frontiers in Microbiology **7**.

64. **Nam Y-D, Lee S-Y, Lim S-I.** 2012. Microbial community analysis of Korean soybean pastes by next-generation sequencing. International Journal of Food Microbiology **155:**36-42.

65. **Kwon DY, Chung KR, Yang H-J, Jang D-J.** 2015. Gochujang (Korean red pepper paste): A Korean ethnic sauce, its role and history. Journal of Ethnic Foods **2:**29-35.

66. **Nam YD, Park Sl, Lim SI.** 2012. Microbial composition of the Korean traditional food “kochujang” analyzed by a massive sequencing technique. Journal of Food Science **77**.

67. **Kim MJ, Kwak HS, Jung HY, Kim SS.** 2016. Microbial communities related to sensory attributes in Korean fermented soy bean paste (doenjang). Food Research International **89:**724-732.

68. **Jung WY, Jung JY, Lee HJ, Jeon CO.** 2016. Functional characterization of bacterial communities responsible for fermentation of Doenjang: a traditional Korean fermented soybean paste. Frontiers in Microbiology **7**.

69. **Kim MJ, Kwak HS, Kim SS.** 2018. Effects of salinity on bacterial communities, Maillard reactions, isoflavone composition, antioxidation and antiproliferation in Korean fermented soybean paste (doenjang). Food Chemistry **245:**402-409.

70. **Yang L, Yang H-l, Tu Z-c, Wang X-l.** 2016. High-Throughput Sequencing of Microbial Community Diversity and Dynamics during Douchi Fermentation. PLoS One **11:**e0168166.

71. **Lee MH, Li FZ, Lee J, Kang J, Lim SI, Nam YD.** 2017. Next-Generation Sequencing Analyses of Bacterial Community Structures in Soybean Pastes Produced in Northeast China. Journal of Food Science **82:**960-968.

72. **Sulaiman J, Gan HM, Yin WF, Chan KG.** 2014. Microbial succession and the functional potential during the fermentation of Chinese soy sauce brine. Front Microbiol **5:**556.

73. **Villarreal-Soto SA, Beaufort S, Bouajila J.** 2018. Understanding Kombucha Tea Fermentation: A Review. Journal of Food Science **83:**580-588.

74. **Vina I, Semjonovs P, Linde R, Denina I.** 2014. Current evidence on physiological activity and expected health effects of kombucha fermented beverage. J Med Food **17:**179-188.

75. **Marsh AJ, O'Sullivan O, Hill C, Ross RP, Cotter PD.** 2014. Sequence-based analysis of the bacterial and fungal compositions of multiple kombucha (tea fungus) samples. Food Microbiology **38:**171-178.

76. **Coton M, Pawtowski A, Taminiau B, Burgaud G, Deniel F, Coulloumme-Labarthe L, Fall A, Daube G, Coton E.** 2017. Unraveling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture-based methods. FEMS Microbiology Ecology **93**.

77. **Reva ON, Zaets IE, Ovcharenko LP, Kukharenko OE, Shpylova SP, Podolich OV, de Vera J-P, Kozyrovska NO.** 2015. Metabarcoding of the kombucha microbial community grown in different microenvironments. AMB Express **5:**35.

78. **De Filippis F, Troise AD, Vitaglione P, Ercolini D.** 2018. Different temperatures select distinctive acetic acid bacteria species and promotes organic acids production during Kombucha tea fermentation. Food Microbiology **73:** 11-16.

79. **Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML.** 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution **4:**1111-1119.

80. **Chakravorty S, Bhattacharya S, Chatzinotas A, Chakraborty W, Bhattacharya D, Gachhui R.** 2016. Kombucha tea fermentation: Microbial

and biochemical dynamics. International Journal of Food Microbiology **220:**63-72.

81. **Zhang L, Zhang Z-z, Zhou Y-b, Ling T-j, Wan X-c.** 2013. Chinese dark teas: Postfermentation, chemistry and biological activities. Food Research International **53:**600-607.

82. **Li Q, Huang J, Li Y, Zhang Y, Luo Y, Chen Y, Lin H, Wang K, Liu Z.** 2017. Fungal community succession and major components change during manufacturing process of Fu brick tea. Scientific Reports **7:**6947.

83. **Mao Y, Wei B, Teng J, Huang L, Xia N.** 2017. Analyses of fungal community by Illumina MiSeq platforms and characterization of Eurotium species on Liupao tea, a distinctive post-fermented tea from China. Food Research International **99:**641-649.

84. **Mo H, Zhu Y, Chen Z.** 2008. Microbial fermented tea – a potential source of natural food preservatives. Trends in Food Science & Technology **19:**124-130.

85. **Lyu C, Chen C, Ge F, Liu D, Zhao S, Chen D.** 2013. A preliminary metagenomic study of puer tea during pile fermentation. Journal of the Science of Food and Agriculture **93:**3165-3174.

86. **Zhang Y, Skaar I, Sulyok M, Liu X, Rao M, Taylor JW.** 2016. The Microbiome and Metabolites in Fermented Pu-erh Tea as Revealed by High-Throughput Sequencing and Quantitative Multiplex Metabolite Analysis. PLoS One **11:**e0157847.

87. **Zhao M, Zhang DL, Su XQ, Duan SM, Wan JQ, Yuan WX, Liu BY, Ma Y, Pan YH.** 2015. An Integrated Metagenomics/Metaproteomics

Investigation of the Microbial Communities and Enzymes in Solid-state Fermentation of Pu-erh tea. Sci Rep **5:**10117.

88. **Ma Y, Duan S, Zhang D, Su X, Zhang D, Lv C, Zhao M.** 2017. Microbial Succession and the Dynamics of Chemical Compounds during the Solid-State Fermentation of Pu-erh Tea. Applied Sciences **7:**166.

89. **Chavan RS, Chavan SR.** 2011. Sourdough Technology—A Traditional Way for Wholesome Foods: A Review. Comprehensive Reviews in Food Science and Food Safety **10:**169-182.

90. **Lattanzi A, Minervini F, Di Cagno R, Diviccaro A, Antonielli L, Cardinali G, Cappelle S, De Angelis M, Gobbetti M.** 2013. The lactic acid bacteria and yeast microbiota of eighteen sourdoughs used for the manufacture of traditional Italian sweet leavened baked goods. International Journal of Food Microbiology **163:**71-79.

91. **Lhomme E, Lattanzi A, Dousset X, Minervini F, De Angelis M, Lacaze G, Onno B, Gobbetti M.** 2015. Lactic acid bacterium and yeast microbiotas of sixteen French traditional sourdoughs. International Journal of Food Microbiology **215:**161-170.

92. **Michel E, Monfort C, Deffrasnes M, Guezenec S, Lhomme E, Barret M, Sicard D, Dousset X, Onno B.** 2016. Characterization of relative abundance of lactic acid bacteria species in French organic sourdough by cultural, qPCR and MiSeq high-throughput sequencing methods. International Journal of Food Microbiology **239:**35-43.

93. **Lhomme E, Orain S, Courcoux P, Onno B, Dousset X.** 2015. The predominance of Lactobacillus sanfranciscensis in French organic

sourdoughs and its impact on related bread characteristics. International Journal of Food Microbiology **213:**40-48.

94. **Alfonzo A, Miceli C, Nasca A, Franciosi E, Ventimiglia G, Di Gerlando R, Tuohy K, Francesca N, Moschetti G, Settanni L.** 2017. Monitoring of wheat lactic acid bacteria from the field until the first step of dough fermentation. Food Microbiology **62:**256-269.

95. **Ercolini D, Pontonio E, De Filippis F, Minervini F, La Storia A, Gobbetti M, Di Cagno R.** 2013. Microbial ecology dynamics during rye and wheat sourdough preparation. Applied and Environmental Microbiology **79:**7827-7836.

96. **Bessmeltseva M, Viiard E, Simm J, Paalme T, Sarand I.** 2014. Evolution of bacterial consortia in spontaneously started rye sourdoughs during two months of daily propagation. PLoS One **9:**e95449.

97. **Rizzello CG, Cavoski I, Turk J, Ercolini D, Nionelli L, Pontonio E, De Angelis M, De Filippis F, Gobbetti M, Di Cagno R.** 2015. Organic cultivation of Triticum turgidum subsp. durum is reflected in the flour-sourdough fermentation-bread axis. Applied and Environmental Microbiology **81:**3192-3204.

98. **Minervini F, Celano G, Lattanzi A, De Angelis M, Gobbetti M.** 2016. Added ingredients affect the microbiota and biochemical characteristics of durum wheat type-I sourdough. Food Microbiology **60:**112-123.

99. **Minervini F, Lattanzi A, De Angelis M, Celano G, Gobbetti M.** 2015. House microbiotas as sources of lactic acid bacteria and yeasts in traditional Italian sourdoughs. Food Microbiology **52:**66-76.

100. **Viiard E, Bessmeltseva M, Simm J, Talve T, Aaspõllu A, Paalme T, Sarand I.** 2016. Diversity and stability of lactic acid bacteria in rye sourdoughs of four bakeries with different propagation parameters. PLoS One **11:**e0148325.

101. **Lee C-H.** 1997. Lactic acid fermented foods and their benefits in Asia. Food Control **8:**259-269.

102. **Roh SW, Kim K-H, Nam Y-D, Chang H-W, Park E-J, Bae J-W.** 2010. Investigation of archaeal and bacterial diversity in fermented seafood using barcoded pyrosequencing. The ISME Journal **4:**1-16.

103. **Zhang J, Wang X, Huo D, Li W, Hu Q, Xu C, Liu S, Li C.** 2016. Metagenomic approach reveals microbial diversity and predictive microbial metabolic pathways in Yucha, a traditional Li fermented food. Scientific Reports **6:**32524.

104. **Kiyohara M, Koyanagi T, Matsui H, Yamamoto K, Take H, Katsuyama Y, Tsuji A, Miyamae H, Kondo T, Nakamura S.** 2012. Changes in microbiota population during fermentation of narezushi as revealed by pyrosequencing analysis. Bioscience, Biotechnology, and Biochemistry **76:**48-52.

105. **Koyanagi T, Kiyohara M, Matsui H, Yamamoto K, Kondo T, Katayama T, Kumagai H.** 2011. Pyrosequencing survey of the microbial diversity of ‘narezushi’, an archetype of modern Japanese sushi. Letters in Applied Microbiology **53:**635-640.

106. **Koyanagi T, Nakagawa A, Kiyohara M, Matsui H, Yamamoto K, Barla F, Take H, Katsuyama Y, Tsuji A, Shijimaya M.** 2013. Pyrosequencing

analysis of microbiota in Kaburazushi, a traditional medieval sushi in Japan. Bioscience, Biotechnology, and Biochemistry **77:**2125-2130.

107. **Kim HJ, Kim M-J, Turner TL, Kim B-S, Song K-M, Yi SH, Lee M-K.** 2014. Pyrosequencing analysis of microbiota reveals that lactic acid bacteria are dominant in Korean flat fish fermented food, gajami-sikhae. Bioscience, Biotechnology, and Biochemistry **78:**1611-1618.

108. **Duan S, Hu X, Li M, Miao J, Du J, Wu R.** 2016. Composition and Metabolic Activities of the Bacterial Community in Shrimp Sauce at the Flavor-Forming Stage of Fermentation As Revealed by Metatranscriptome and 16S rRNA Gene Sequencings. Journal of Agricultural and Food Chemistry **64:**2591-2603.

109. **Ojha KS, Kerry JP, Duffy G, Beresford T, Tiwari BK.** 2015. Technological advances for enhancing quality and safety of fermented meat products. Trends in Food Science & Technology **44:**105-116.

110. **Toldrá F.** 2011. 20 - Improving the sensory quality of cured and fermented meat products, p 508-526, Processed Meats doi:https://doi.org/10.1533/9780857092946.3.508. Woodhead Publishing.

111. **Greppi A, Ferrocino I, La Storia A, Rantsiou K, Ercolini D, Cocolin L.** 2015. Monitoring of the microbiota of fermented sausages by culture independent rRNA-based approaches. International Journal of Food Microbiology **212:**67-75.

112. **Połka J, Rebecchi A, Pisacane V, Morelli L, Puglisi E.** 2015. Bacterial diversity in typical Italian salami at different ripening stages as revealed by high-throughput sequencing of 16S rRNA amplicons. Food Microbiology **46:**342-356.

113. **Fontana C, Bassi D, López C, Pisacane V, Otero MC, Puglisi E, Rebecchi A, Cocconcelli PS, Vignolo G.** 2016. Microbial ecology involved in the ripening of naturally fermented llama meat sausages. A focus on lactobacilli diversity. International Journal of Food Microbiology **236:**17-25.

114. **Wang X, Ren H, Zhan Y.** 2017. Characterization of microbial community composition and pathogens risk assessment in typical Italian-style salami by high-throughput sequencing technology. Food Science and Biotechnology doi:10.1007/s10068-017-0200-5.

115. **Quijada NM, De Filippis F, Sanz JJ, del Camino García-Fernández M, Rodríguez-Lázaro D, Ercolini D, Hernández M.** 2018. Different Lactobacillus populations dominate in “Chorizo de León” manufacturing performed in different production plants. Food Microbiology **70:**94-102.

116. **Wang X, Zhang Y, Ren H, Zhan Y.** 2018. Comparison of bacterial diversity profiles and microbial safety assessment of salami, Chinese dry-cured sausage and Chinese smoked-cured sausage by high-throughput sequencing. LWT **90:**108-115.

117. **Ferrocino I, Bellio A, Giordano M, Macori G, Romano A, Rantsiou K, Decastelli L, Cocolin L.** 2018. Shotgun metagenomics and volatilome profile of the microbiota of fermented sausages. Applied and Environmental Microbiology **84:**e02120-02117.

118. **Bamforth CW.** 2017. Progress in Brewing Science and Beer Production. Annu Rev Chem Biomol Eng **8:**161-176.

119. **Morgan HH, du Toit M, Setati ME.** 2017. The Grapevine and Wine Microbiome: Insights from High-Throughput Amplicon Sequencing. Frontiers in Microbiology **8**.

120. **Zou W, Zhao C, Luo H.** 2018. Diversity and Function of Microbial Community in Chinese Strong-Flavor Baijiu Ecosystem: A Review. Frontiers in Microbiology **9**.

121. **Bokulich NA, Collins TS, Masarweh C, Allen G, Heymann H, Ebeler SE, Mills DA.** 2016. Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. mBio **7:**e00631-00616.

122. **Stefanini I, Carlin S, Tocci N, Albanese D, Donati C, Franceschi P, Paris M, Zenato A, Tempesta S, Bronzato A.** 2017. Core microbiota and metabolome of Vitis vinifera L. cv. Corvina Grapes and Musts. Frontiers in Microbiology **8**.

123. **Varela C, Barker A, Tran T, Borneman A, Curtin C.** 2017. Sensory profile and volatile aroma composition of reduced alcohol Merlot wines fermented with Metschnikowia pulcherrima and Saccharomyces uvarum. International Journal of Food Microbiology **252:**1-9.

124. **Hong X, Chen J, Liu L, Wu H, Tan H, Xie G, Xu Q, Zou H, Yu W, Wang L.** 2016. Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. Scientific Reports **6:**26621.

125. **Wang X, Du H, Zhang Y, Xu Y.** 2017. Environmental Microbiota Drives Microbial Succession and Metabolic Profiles during Chinese Liquor Fermentation. Applied and Environmental Microbiology**:**AEM. 02369-02317.

126. **Song Z, Du H, Zhang Y, Xu Y.** 2017. Unraveling core functional microbiota in traditional solid-state fermentation by high-throughput

amplicons and metatranscriptomics sequencing. Frontiers in Microbiology **8:**1294.

127. **Liu J, Wu Q, Wang P, Lin J, Huang L, Xu Y.** 2017. Synergistic effect in core microbiota associated with sulfur metabolism in spontaneous Chinese liquor fermentation. Applied and Environmental Microbiology **83:**e01475-01417.

128. **Li S, Li P, Feng F, Luo L-X.** 2015. Microbial diversity and their roles in the vinegar fermentation process. Applied Microbiology and Biotechnology **99:**4997-5024.

129. **Wang Z-M, Lu Z-M, Yu Y-J, Li G-Q, Shi J-S, Xu Z-H.** 2015. Batch-to-batch uniformity of bacterial community succession and flavor formation in the fermentation of Zhenjiang aromatic vinegar. Food Microbiology **50:**64-69.

130. **Nie Z, Zheng Y, Wang M, Han Y, Wang Y, Luo J, Niu D.** 2013. Exploring microbial succession and diversity during solid-state fermentation of Tianjin duliu mature vinegar. Bioresource technology **148:**325-333.

131. **Li S, Li P, Liu X, Luo L, Lin W.** 2016. Bacterial dynamics and metabolite changes in solid-state acetic acid fermentation of Shanxi aged vinegar. Applied Microbiology and Biotechnology **100:**4395-4411.

132. **Wang Z-M, Lu Z-M, Shi J-S, Xu Z-H.** 2016. Exploring flavour-producing core microbiota in multispecies solid-state fermentation of traditional Chinese vinegar. Scientific Reports **6:**26818.

133. **Lu Z-M, Liu N, Wang L-J, Wu L-H, Gong J-S, Yu Y-J, Li G-Q, Shi J-S, Xu Z-H.** 2016. Elucidating and regulating the acetoin production role of

microbial functional groups in multispecies acetic acid fermentation. Applied and Environmental Microbiology **82:**5860-5868.

134. **Segata N.** 2018. On the Road to Strain-Resolved Comparative Metagenomics. mSystems **3**.

135. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

136. **Albanese D, Donati C.** 2017. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun **8:**2260.

137. **Magnúsdóttir S, Thiele I.** 2018. Modeling metabolism of the human gut microbiome. Current Opinion in Biotechnology **51:**90-96.

138. **Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, de Wouters T, Juste C, Rizkalla S, Chilloux J, Hoyles L, Nicholson Jeremy K, Dore J, Dumas Marc E, Clement K, Bäckhed F, Nielsen J.** 2015. Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. Cell Metabolism **22:**320-331.

Chapter 3

Microbial succession and flavour production in the fermented dairy beverage kefir

Figures updated since publication in *mSystems*

(doi: https://doi.org/10.1128/mSystems.00052-16)

Authors: Aaron M. Walsh, Fiona Crispie, Kieran Kilcawley, Orla O'Sullivan, Maurice G. O'Sullivan, Marcus J. Claesson, and Paul D. Cotter.

Contributions:

- **Candidate** performed fermentations, sample collection, DNA extractions, sequencing library preparations, and bioinformatic and statistical analysis
- **KK** performed GC-MS analysis
- **OOS** provided guidance for bioinformatic analysis
- **MOS** performed sensory analysis
- **FC**, **MJC**, and **PDC** supervised the study

ABSTRACT

Kefir is a putatively health-promoting dairy beverage that is produced when a kefir grain, consisting of a consortium of microorganisms, is added to milk to initiate a natural fermentation. Here, a detailed analysis was carried out to determine how the microbial population, gene content and flavour of three kefirs from distinct geographical locations change over the course of 24-hour-fermentations. Metagenomic sequencing revealed that *Lactobacillus kefiranofaciens* was the dominant bacterial species in kefir during early stages of fermentations, but that *Leuconostoc mesenteroides* became more prevalent in later stages. This pattern is consistent with an observation that genes involved in aromatic amino acid biosynthesis were absent from *Lb. kefiranofaciens,* but were present in *L. mesenteroides*. Additionally, these shifts in the microbial community structure, and associated pathways, corresponded to changes in the levels of volatile compounds. Specifically, *Acetobacter* spp. correlated with acetic acid, *Lactobacillus* spp. correlated with carboxylic acids, esters and ketones, *Leuconostoc* spp. correlated with acetic acid and 2,3-butanedione, and *Saccharomyces* spp. correlated with esters. The correlation data suggest a causal relationship between microbial taxa and flavour which is supported by observations that addition of *Lb. kefiranofaciens* NCFB 2797 increased the levels of esters and ketones, whereas addition of *L. mesenteroides* DPC 7047 increased acetic acid and 2,3-butanedione. Finally, we detected genes that were potentially associated with probiotic traits, such as bile tolerance or bacteriocin production, in the kefir microbiome. Our results illustrate the dynamic nature of kefir fermentations and microbial succession patterns therein, and can be applied to optimise fermentation processes, flavours and health-related attributes of this and other fermented foods.

INTRODUCTION

Our knowledge of the composition of complex microbial communities from different environments has increased dramatically in recent years (1-3). However, considerably less is known about the biological interactions and other processes which drive microbial succession, or changes in the microbial population structure over time, in these environments (4). It has been proposed that microbial communities from fermented foods could provide a useful model for elucidating the determinants of microbial succession, given that they are considerably less complex than, for example, those from the gut or soil (5). Indeed, cheese rind communities have previously been used to great effect for this purpose (6).

Here, we show that kefir provides an alternative model microbial community that is less complex and provides results even more quickly. Kefir is a traditional fermented milk beverage that is typically produced by inoculating a kefir grain, a cauliflower-like exopolysaccharide matrix containing a symbiotic community of bacteria and yeast (7), into milk and incubating it at room temperature for approximately 24 hours resulting in a beverage that has been described as having a pleasantly sour or yoghurt-like taste (8). This flavour can vary depending on the microbial composition of the grain that is used (9). High-throughput sequencing investigations have demonstrated that kefir grains are typically dominated by the bacterial genus *Lactobacillus* and the fungal phylum Ascomycota (9, 10). In contrast, kefir milk is dominated by the bacterial genera *Lactobacillus*, *Lactococcus*, *Acetobacter*, and *Leuconostoc*, and the fungal genera *Kazachstania*, *Kluyveromyces*, *Naumovozyma*, and *Saccharomyces* (9, 11, 12).

The consumption of kefir has been associated with numerous health benefits, including anti-carcinogenic, anti-inflammatory and anti-pathogenic effects (13-15), as well as the alleviation of the symptoms of lactose intolerance and the reduction of cholesterol (16, 17). There is mounting evidence to suggest that the microorganisms present in kefir exert at least some of these health benefits (18-22) but there is a lack of understanding of the mechanisms by which they do so.

In this work, amplicon sequencing and whole metagenome shotgun sequencing are combined with metabolomics and flavour analysis to highlight how microbial composition, gene content and flavour of kefir change over the course of 24-hour fermentations. We demonstrate that the integration of multi-omics data can predict the contribution of individual microorganisms to metabolite production in a microbial environment, using flavour formation as an example, and we validate these findings through supplementation with specific microbes. To our knowledge, this is the first study to combine metagenome binning and metabolic reconstruction to determine the microbial composition, at both species-level and strain-level, and the functional potential of a fermented food, respectively, at different stages of fermentation. In addition, this is the first study to combine whole metagenome shotgun sequencing with metabolomics to link microbial species with volatile production in kefir. Our findings reveal a dynamic flux from *Lactobacillus kefiranofaciens* domination during the early stages of fermentations to *Leuconostoc mesenteroides* domination during the latter stages, establish a causal relationship between microbial taxa and flavour, and highlight genes that likely contribute to kefir's purported health-associated attributes.

MATERIALS AND METHODS

Kefir fermentations

Three kefir grains, Fr1, Ick, and UK3, from distinct geographical locations, France, Ireland and the UK, respectively, were used for kefir fermentations. The grains were weighed and inoculated in full-fat pasteurised milk at a concentration of 2% (w/v) in separate fermentation vessels. The milk was incubated at 25°C for 24 hours. 20 ml of milk was collected after 0, 8 or 24 hours. In total, there were 15 2% (w/v) kefir milk samples: three 0 hour samples that were collected immediately before the addition of Fr1, Ick or UK3, three 8 hour samples (one each from Fr1, Ick and UK3) and nine 24 hour samples (one from each of the three replicate fermentations with Fr1, Ick or UK3). The samples were stored at -20°C until DNA extraction and volatile analysis. Kefir grains were washed with sterile deionised water between fermentations.

Additional fermentations were performed in which milk inoculated with specific kefir grains was supplemented with kefir isolates to assess the consequences of increased levels of these taxa on volatile levels and flavour. Specifically, *Lb. kefiranofaciens* NCFB 2797 and *L. mesenteroides* DPC 7047 were grown overnight in 10 ml of MRS broth, were pelleted at 5,444 *x g* and resuspended in 5 ml pasteurised milk. *Lb. kefiranofaciens* NCFB 2797 cells were added to Fr1 milk and *L. mesenteroides* DPC 7047 cells were added to Ick milk. Non-spiked Fr1 and Ick served as negative controls. As above, milk was incubated at 25°C for 24 hours and the fermentations were carried out in triplicate. 5 ml of milk was collected for volatile analysis and the samples were stored at -20°C. 400 ml of milk was collected for sensory evaluation and the samples were stored at -80°C.

Volatile profiling of kefir by GCMS

For volatile analysis of kefir, 1 g of the sample was added to 20 ml screw capped SPME vial with a silicone/PTFE septum (Apex Scienfific, Maynooth, Ireland) and equilibrated to 75°C for 5 mins with pulsed agitation of 5 seconds at 400 rpm using a GC Sampler 80 (Agilent Technologies Ltd, Little Island, Cork, Ireland). A single 50/30 µm CarboxenTM/divinylbenzene/polydimethylsiloxane (DVB/CAR/PDMS) SPME fiber (Agilent Technologies Ltd, Ireland) was used and was exposed to the headspace above the samples for 20 min at depth of 1 cm at 75°C. The fibre was retracted and injected into the GC inlet and desorbed for 2 min at 250°C. After injection the fibre was heated in a bakeout station for 3 min at 270°C to cleanse the fibre. The samples were analysed in triplicate. Injections were made on an Agilent 7890A GC with an Agilent DB-5 (60 m x 0.25 mm x 0.25 µm) column using a multipurpose injector with a merlin microseal (Agilent Technologies Ltd, Ireland). The temperature of the column oven was set at 35°C, held for 0.5 min, increased at 6.5°C min^{-1} to 230°C then increased at 15°C min^{-1} to 325°C, yielding at total run time of 36.8 min. The carrier gas was helium held at a constant pressure of 23 psi. The detector was an Agilent 5975C MSD single quadrupole mass spectrometer detector (Agilent Technologies Ltd, Ireland). The ion source temperature was 230°C and the interface temperature were set at 280°C and the MS mode was electronic ionization (-70v) with the mass range scanned between 35 and 250 amu. Compounds were identified using mass spectra comparisons to the National Institute of Standards and Technology (NIST) 2011 mass spectral library, Automated Mass Spectral Deconvolution and Identification System (AMDIS) and in-house library created in TargetView software (Markes International, Llantrisant, UK) with target and qualifier ions and linear retention indices for each compound. An auto-tune of

the GCMS was carried out prior to the analysis to ensure optimal GCMS performance. A set of external standards was also run at the start and end of the sample set and abundances were compared to known amounts to ensure that both the SPME extraction and MS detection was performing within specification.

Volatile profiling of spiked and non-spiked kefir samples was done using a slightly modified GCMS protocol, as detailed in Supplemental Materials and Methods.

Sensory analysis of spiked and non-spiked kefir

25 naïve assessors were recruited for sensory acceptance evaluation and 10 trained assessors were recruited for ranking descriptive analysis (RDA). ANOVA-Partial Least Squares regression (APLSR) was used to process the results of the sensory acceptance evaluation test and RDA, using Unscrambler software version10.3. See Supplemental Materials and Methods for a more in depth description of the sensory analysis methods.

Total DNA extraction from kefir (milks and grains)

DNA was extracted from 15 ml of kefir milk as follows: milk was centrifuged at 5,444 *xg* for 30 minutes at 4°C to pellet the microbial cells in the liquid. The cell pellet was resuspended in 200 µl of PowerBead solution from the PowerSoil DNA Isolation Kit (Cambio, Cambridge, UK). The resuspended cells were transferred to a PowerBead tube (Cambio, Cambridge, UK). 90 µl of 50 mg/ml lysozyme (Sigma-Aldrich, Dublin, Ireland) and 50 µl of 100 U/ml mutanolysin (Sigma-Aldrich, Dublin, Ireland) were added and the sample was incubated at 60°C for 15 minutes.

28 µl of Proteinase K (Sigma-Aldrich, Dublin, Ireland) was added and the sample was incubated at 60°C for a further 15 minutes. DNA was then purified from the sample using the standard PowerSoil DNA Isolation Kit protocol (Cambio, Cambridge, UK). Total DNA was also extracted from each of the three grains. 50 mg fragments were removed from different sites on each of the grains and added to separate PowerBead tubes (Cambio, Cambridge, UK). The grain fragments were homogenised by shaking the PowerBead tube on the TissueLyser II (Qiagen, West Sussex, UK) at 20 Hz for 10 minutes. Following homogenisation, DNA was purified from the sample using the method outlined above. Total DNA was initially quantified and qualified using gel electrophoresis and the Nanodrop 1000 (BioSciences, Dublin, Ireland), before more accurate quantification using the Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland). Bacterial and fungal abundances were determined by qPCR using the protocol described by Fouhy *et al.* (23) and the Femto Fungal DNA Quantification Kit (Cambridge Biosciences, UK), respectively.

Amplicon sequencing

16S rRNA gene libraries were prepared from extracted DNA using the 16S Metagenomic Sequencing Library Preparation protocol from Illumina (24). ITS gene libraries were prepared for the samples using a modified version of the 16S rRNA gene extraction protocol; briefly, the initial gDNA amplification was performed with primers specific to the ITS1-ITS2 region of the ITS gene (25), but which were modified to incorporate the Illumina overhang adaptor (i.e. ITSF1 primer 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGGTCATTTAGAGG

AAGTAA-3'; ITS2 primer 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGCGTTCTTCATCGATGC-3'). After amplification of the ITS1-ITS2 region, PCR products were treated as described in the Illumina protocol. Samples were sequenced on the Illumina MiSeq in the Teagasc sequencing facility, using a 2 x 250 cycle V2 kit, following standard Illumina sequencing protocols.

Whole metagenome shotgun sequencing

Whole metagenome shotgun libraries were prepared as per the Nextera XT DNA Library Preparation Guide from Illumina (24). Samples were sequenced on the Illumina MiSeq sequencing platform in the Teagasc sequencing facility, using a 2 x 300 cycle V3 kit, following standard Illumina sequencing protocols.

Bioinformatic analysis

16S rRNA gene sequencing data was processed using the pipeline described by Fouhy *et al.* (26); briefly sequences were quality checked, clustered into operational taxonimcal units (OTUs), aligned and diversity calculated (both alpha and beta) using a combination of the Qiime (1.8.0) (27) and USearch (v7-64bit) (28) pipelines. Taxonomy was assigned using a BLAST (29) against the SILVA SSURef database release 1 (30). ITS gene sequencing data was processed using a slightly modified pipeline: taxonomy was assigned using BLAST against the ITSoneDB database (31). Raw reads from whole metagenome shotgun sequencing were filtered based on quality and quantity and were trimmed to 200 bp with a combination of Picardtools

(http://broadinstitute.github.io/picard/) and SAMtools (32). Subsequently function was assigned to reads using the HUMAnN2 suite of tools (33), which assigned function based on the ChocoPhlan databases and genes based on UniRef (34). The HUMAnN 2 gene abundance table was regrouped using a mapping of MetaCyc pathways and a mapping of Gene Ontology (GO) terms for amino acid, carbohydrate and lipid metabolism. MetaPhlAn2 and Kraken were used to profile changes in the microbial composition of kefir milk at the species level (35, 36).

Statistical analysis of metagenomic and metabolomic data

Statistical analysis was done using R-3.2.2 (37) and LEfSe (38). The R packages ggplot2 and gplots, and the cladogram generator Graphlan (39) were used for data visualisation.

RESULTS

Microbial composition of kefir

16S rRNA and ITS gene sequencing were used to determine the changes in the microbial population of kefirs over the course of 24-hour fermentations initiated with three separate grains, designated Fr1, Ick and UK3, from distinct geographic locations, namely France, Ireland and the United Kingdom.

Analysis of the grains showed that *Lactobacillus* was the dominant bacterial genus and constituted >92% of the population of all three grains (Figure S1). *Acetobacter* was subdominant, and accounted for between 1 to 2% of the population of each

grain. In addition, *Leuconostoc* was present in all three grains, although its abundance varied from 0.2 to 1.5%. Other genera that were detected at a relative abundance >1% were *Propionibacterium,* in Fr1 (4.6%) only, and *Bifidobacterium,* in UK3 (3.4%) only. A fungal population was detected in the grains Fr1 and Ick, but not in UK3. *Saccharomyces* and *Kazachstania* were the only fungal genera present (Figure S1).

Analysis of milk samples revealed that an initially relatively high bacterial diversity decreased over time, with a small number of genera becoming dominant by 8 and 24 hours (Figure S2). On average, at 0 hours, or immediately before the grains were added to the milk, the bacterial genera present at a relative abundance ≥1% were *Pseudomonas* (16.9%), *Anoxybacillus* (7.1%), *Thermus* (6.5%), *Acinetobacter* (5%), *Streptococcus* (4.5%), *Geobacillus* (3.2%), *Clostridium* (2.4%), *Butyrivibrio* (2.2%), *Serratia* (2.1%), *Enterobacter* (1.3%), *Turicibacter* (1.3%), and *Lactococcus* (1%). A further 46.5% of bacterial genera had a relative abundance <1% (Figure 1.A). This microbial profile is consistent with that of pasteurised milk as reported previously by Quigley *et al*. (40). We were unable to generate an ITS amplicon for the three samples collected at 0 hours, and quantitative PCR (qPCR) indicated fungal DNA was present at less than 2 pg/µl.

qPCR measurements revealed that total bacterial and fungal levels increased after kefir grains were added to milk (Table S1). At 8 and 24 hours in Fr1, Ick and UK3, *Lactobacillus, Leuconostoc* and *Acetobacter* accounted for >98% of the total bacterial population, while *Saccharomyces* and *Kazachstania* accounted for over 99% of the fungal population. No other bacterial or fungal genera were present at a relative abundance >1%.

Although there were some differences in their composition at each time-point, the bacterial communities of the three kefirs all followed the same pattern of succession (Figure 1.A). Between 0 and 8 hours, there was an increase in the relative abundances of *Lactobacillus*, *Leuconostoc* and *Acetobacter*. *Lactobacillus* was the dominant genus at 8 hours. However, between 8 and 24 hours, the relative abundance of *Lactobacillus* decreased. Concurrently, the relative abundances of *Leuconostoc* and *Acetobacter* increased. On average, *Leuconostoc* accounted for approximately one-third of the bacterial population at 24 hours. In contrast to the bacterial communities of the three kefirs, the respective fungal communities displayed varying patterns of succession (Figure S3.A).

16S rRNA and ITS compositional data were supplemented by composition-based analysis of shotgun metagenomics data. Kraken (36) was used to determine the bacterial composition of kefir after 0, 8 and 24 hours of fermentation and yielded results that corresponded well with amplicon sequencing results at the genus level, but which could be further assigned to the species level. It was established that the kefir milk was dominated by *Lb. kefiranofaciens* at 8 hours (Figure 1.B). However, between 8 and 24 hours, the relative abundance of *Lb. kefiranofaciens* decreased, whereas the relative abundance of *Leuconostoc mesenteroides* increased. During the same period there were also increases in the relative abundances of *Acetobacter pasteurianus*, *Lactobacillus helveticus*, *Leuconostoc citreum*, *Leuconostoc gelidum*, and *Leuconostoc kimchii*. These results were generally consistent with those generated by MetaPhlan2 (35) (Figure S3.B), except that MetaPhlan2 did not detect some of the species present in lower abundance (i.e. *A. pasteurianus*, *L. citreum*, *L. gelidum* or *L. kimchii*). MetaPhlan2 predicted that *Saccharomyces cerevisiae* was the dominant fungal species, and that it accounted for 0.9% and 0.2% of the microbiota

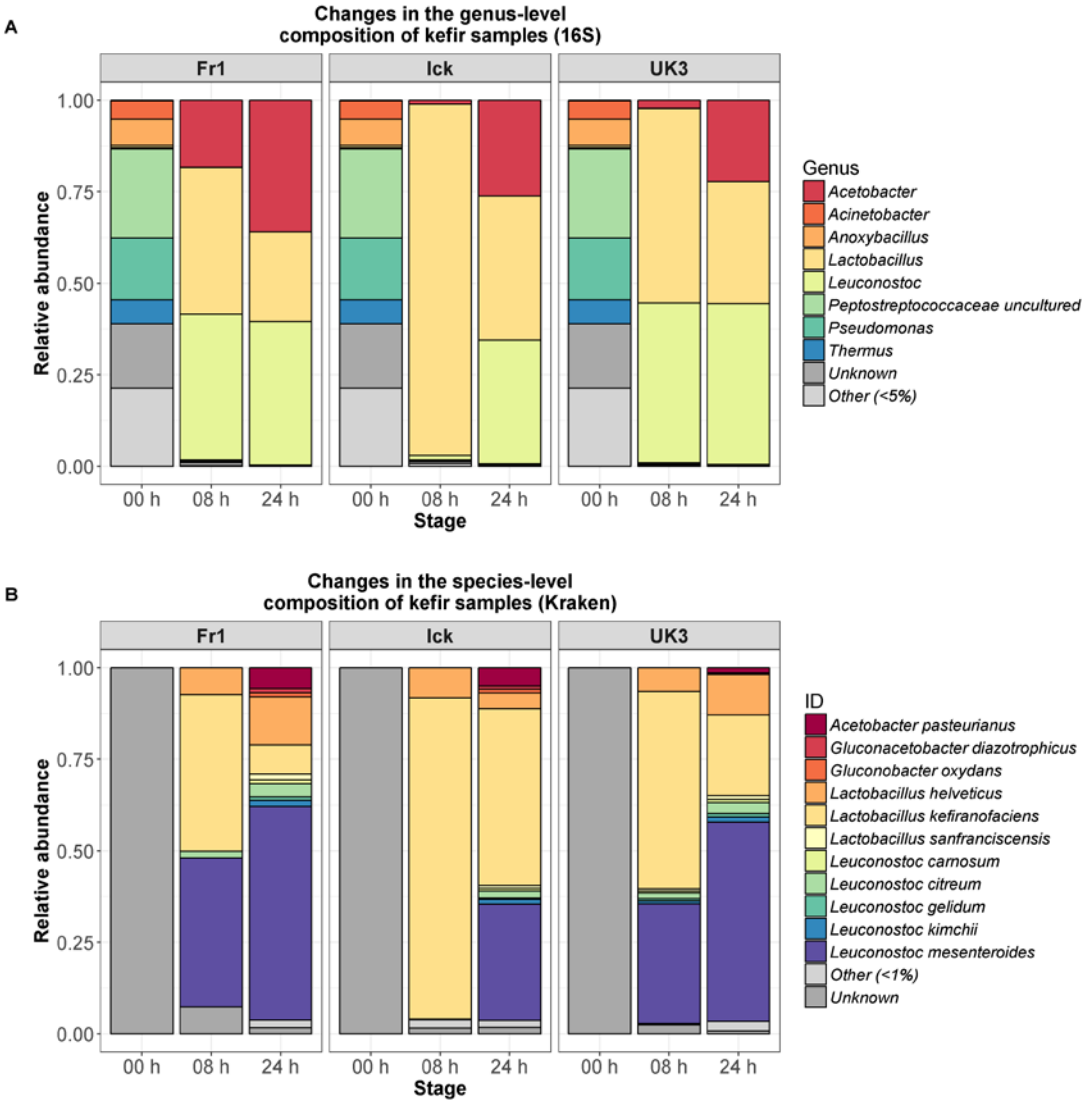


Figure 1: Stacked bar charts presenting the bacterial composition of kefir samples after 0, 8 and 24 hours of fermentation, as determined by (a) 16S rRNA gene sequencing, and (b) binning of metagenome sequences using Kraken.

in kefir at 8 and 24 hours of fermentation, respectively. However, it did not detect *Kazachstania* species.

In addition, PanPhlAn (41) was used to provide strain level characterisation of the most dominant bacterial species identified by Kraken and MetaPhlan2. The results indicated that, across all kefirs, the strains present were most closely related to *Lb. kefiranofaciens* DSM 10550, *L. mesenteroides* ATCC 8293 and *L. helveticus* MTCC 5463 (Figure S5). Despite this relative homogeneity, it was still apparent that the strains in a particular kefir were more closely related to each other than they were to strains from other kefirs (Figure S4).

Gene content of kefir

Whole metagenome shotgun sequencing was used to characterise the functional potential of the kefir microbiome at different stages of fermentation and the HUMAnN2 pipeline (https://bitbucket.org/biobakery/humann2) was used for metagenomic metabolic reconstruction. The default HUMAnN2 pathway abundance table was regrouped using a custom mapping file to assign individual MetaCyc pathways (42) to a hierarchy of 534 gene product categories to achieve an overview of the kefir microbiome (Figure 2). The statistical tool LEfSe (38) was used to identify changes in the abundances of genetic pathways over the course of fermentation. Notably, we observed that pathways involved in carbohydrate metabolism, carboxylate degradation and unsaturated fatty acid biosynthesis were most prevalent at 8 hours, whereas those involved in amino acid metabolism and 2,3-butanediol degradation were most prevalent at 24 hours (Figure 2). Inspection of the default pathway abundance table revealed that pathways involved in fatty acid beta-

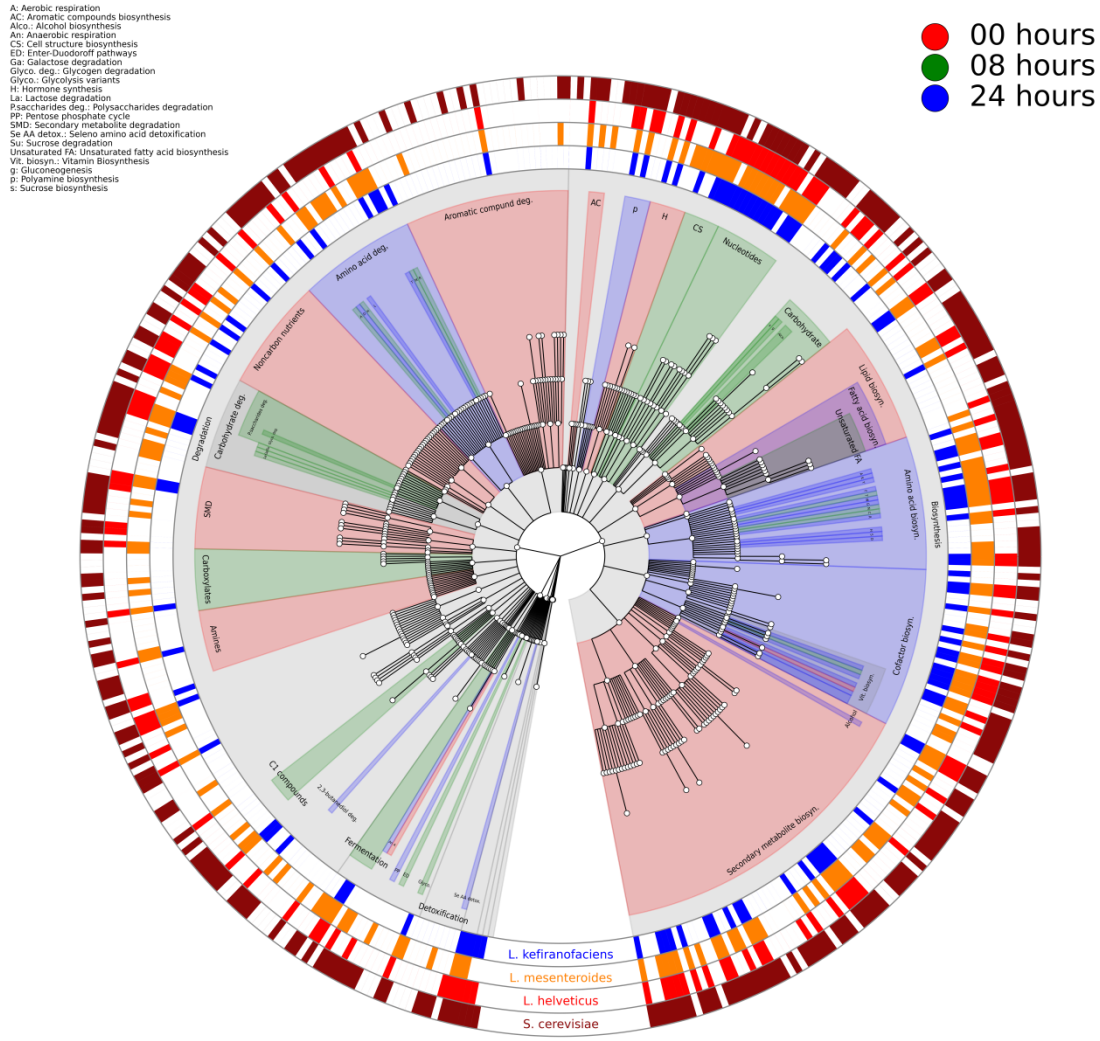


Figure 2: Cladogram presenting a hierarchical overview of the MetaCyc pathways detected in the kefir microbiome using HUMAnN2. Central nodes represent general pathway category functions, like carbohydrate catabolism, and their descendant nodes represent more specific pathway category functions, like sucrose degradation. The colours of the clades indicate the time at which pathways of particular interest were most prevalent, as determined by LEfSe. The outer rings indicate the presence/absence of pathways in *Lb. kefiranofaciens* (blue), *L. mesenteroides* (orange), *L. helveticus* (red), and *S. cerevisiae* (maroon).

oxidation were present in kefirs. The pathways mentioned here are of particular interest because they are potentially involved in the production of volatile compounds (Table 1).

In addition, the default HUMAnN2 gene families table was regrouped to Gene Ontology (GO) terms (gene product categories (43)) and, in total, we detected 1,288, 1,006 and 947 GO terms associated with carbohydrate, amino acid and lipid metabolism in the kefir microbiome. Interestingly, pathways involved in aromatic amino acids and proline biosynthesis were assigned to *L. mesenteroides*, but not *Lb. kefiranofaciens.* Similarly, pathways involved in arabinose, maltose, pentose, sucrose, xylose and xylulose metabolism were present in *L. mesenteroides* but not in *Lb. kefiranofaciens.*

Finally, the HUMANnN2 gene families table was inspected for genes associated with probiotic functionalities to better understand the basis of the health benefits of kefir. We observed that *Lb. kefiranofaciens* in Fr1, Ick and UK3 contained genes encoding exopolysaccharide (EPS) synthesis proteins (UniRef50_W5XGS2, UniRef50_F6CC46 and UniRef50_F0TGY1), bile salt transporter proteins (UniRef50_Q74LX5 and UniRef50_F6CE74), adhesion proteins (UniRef50_F6CFB4 and UniRef50_Q040W2), mucus binding proteins (UniRef50_F6CE70, UniRef50_F6CE69, UniRef50_F6CDG7 and UniRef50_F6CBX6), and the type III bacteriocins/bacteriolysins helveticin J (UniRef50_D5GYX2) and enterolysin A (UniRef50_D5GXY3 and UniRef50_F6CAP6). On the basis of these findings, we downloaded publicly available metagenome sequences from cheeses and kimchi (Table 2) to determine the prevalence of similar genes in other fermented foods. HUMAnN2 indicated that genes encoding EPS synthesis proteins, adhesion proteins,

Table 1: Volatile compounds detected in kefir using GC-MS.

Compound	LRI[a]	Ref LRI[b]	Odour descriptor	Source
Carboxylic acids:				
Acetic acid	692	629	*Vinegar, peppers, green, fruity floral, sour*	Carbohydrate metabolism
Hexanoic acid	968	983	*Sweaty, cheesey, sharp, goaty, bad breath, acidic*	Lipid metabolism
Octanoic acid	1163	1160	*Cheesey, rancid, pungent, sweat, soapy, goaty*	Lipid metabolism
Nonanoic acid	1254	1276	*Fatty, soapy, waxy, green, goat*	Lipid metabolism
n-Decanoic acid	1355	1379	*Soapy, waxy, stale, buttery, fruity, grassy, cheesey, milky*	Lipid metabolism
Alcohols:				
2-Methyl-1-butanol	733	755	*Penetrating, alcohol, wine-like, plastic*	Amino acid metabolism
2-Ethyl-1-hexanol	1025	1031	*Animal, cardboard*	Lipid metabolism
Ethanol	468	426	*Dry, dust*	Carbohydrate metabolism
2-Butanol	601	596	*Fruity*	Carbohydrate metabolism
2-Methyl-1-propanol	621	647	*Malty*	Amino acid metabolism
3-Methyl-Butanol	730	768	*Fresh cheese, breathtaking, alcoholic, fruity, grainy, solvent-like, floral, malty*	Amino acid metabolism
Phenylethyl alcohol	1119	1112	*Unclean, rose, violet-like, honey, floral, spicy*	Amino acid metabolism
1-Pentanol	730	768	*Fruity, alcoholic, green, balsamic, fusel oil, woody*	Lipid metabolism
Aldehydes:				
3-Methyl-butanal	649	654	*Malty, cheesey, green, dark chocolate, cocoa*	Amino acid metabolism
2-Methyl-butanal	658	662	*Malty, dark chocolate, almond, cocoa, coffee*	Amino acid metabolism
Octanal	1002	1004	*Green, fatty, soapy, fruity, orange peel*	Lipid metabolism
Nonanal	1103	1106	*Green, citrus, fatty, floral*	Lipid metabolism
Pentanal	694	697	*Pungent, almond-like, chemical, malty, apple, green*	Lipid metabolism
Hexanal	798	801	*Green, slightly fruity, lemon, herbal, grassy, tallow*	Lipid metabolism
Heptanal	900	901	*Slightly fruity (Balsam), fatty, oily, green, woody*	Lipid metabolism
Esters:				
Ethyl acetate	609	614	*Solvent, pineapple, fruity, apples*	Carbohydrate metabolism
Ethyl butanoate	802	800	*Ripe fruit, buttery, green, apple, pineapple, banana, sweet*	Carbohydrate metabolism
Ethyl hexanoate	995	1002	*Fruity, malty, young cheese, mouldy, apple, green, orange, pineapple, banana*	Carbohydrate metabolism
Ethyl octanoate	1190	1198	*Fruity, apple, green, fatty, orange, winey, pineapple, apricot*	Carbohydrate metabolism
Ethyl decanoate	1388	1396	*Fruity, grape, cognac*	Carbohydrate metabolism
3-Methyl-1-butanol, acetate	874	879	*Fruity, bannana, candy, sweet; apple peel*	Unknown
Ketones:				
2,3-Butanedione	589	596	*Buttery, strong*	Carbohydrate metabolism
2,3-Pentanedione	694	693	*Creamy, cheesey, oily, sweet buttery, carmellic*	Carbohydrate metabolism
2,3-Hexanedione	781	788	*Sweet, creamy, caramellic, buttery*	Carbohydrate metabolism
2-Heptanone	887	891	*Blue cheese, spicy, roquefort*	Lipid metabolism
2-Undecanone	1288	1294	*Floral, fruity, green, musty, tallow*	Lipid metabolism
2-Pentanone	679	687	*Orange peel, sweet, fruity*	Lipid metabolism
2-Nonanone	1088	1094	*Malty, fruity, hot milk, smoked cheese*	Lipid metabolism
Acetone	494	496	*Earthy, fruity, wood pulp, hay*	Lipid metabolism
2-Butanone	598	593	*Buttery, sour milk, etheric*	Carbohydrate metabolism
Sulphur compounds:				
Dimethyl sulfone	920	926	*Sulphurous, hot milk, burnt*	Amino acid metabolism
Carbon disulfide	537	568	*Sweet, ethereal*	Amino acid metabolism

Table 2: Accession numbers of the cheese and kimhci metagenomes analysed in this study.

Origin	Repository	Accession number	Sample description	Reference
Cheese	MG-RAST	4524483.3	Washed unpasteurised cow's cheese	Wolfe et al., 2014
		4524484.3	Bloomy unpasteurised goat's cheese	
		4524488.3	Natural unpasteurised cow's cheese	
		4524489.3	Bloomy unpasteurised goat's cheese	
		4524490.3	Natural unpasteurised cow's cheese	
		4524493.3	Natural unpasteurised cow's cheese	
		4524494.3	Washed pasteurised cow's cheese	
		4524495.3	Washed unpasteurised cow's cheese	
		4524496.3	Washed unpasteurised cow's cheese	
		4524497.3	Natural pasteurised cow's cheese	
		4524499.3	Washed unpasteurised cow's cheese	
		4524500.3	Washed pasteurised cow's cheese	
		4524505.3	Washed unpasteurised cow's cheese	
	European Nucleotide archive	SAMEA3232870	Continental-type cheese	Quigley et al., 2016
		SAMEA3232871		
		SAMEA3232872		
		SAMEA3232873		
		SAMEA3232874		
		SAMEA3232875		
		SAMEA3232876		
		SAMEA3232877		
		SAMEA3232878		
		SAMEA3232879		
Kimchi	Short Read Archive	SRX072929	Kimchi fermentation: Day 1	Jung et al., 2011
		SRX072930	Kimchi fermentation: Day 7	
		SRX072931	Kimchi fermentation: Day 13	
		SRX072932	Kimchi fermentation: Day 16	
		SRX072933	Kimchi fermentation: Day 18	
		SRX072934	Kimchi fermentation: Day 21	
		SRX072935	Kimchi fermentation: Day 23	
		SRX072936	Kimchi fermentation: Day 25	
		SRX072937	Kimchi fermentation: Day 27	
		SRX072938	Kimchi fermentation: Day 29	

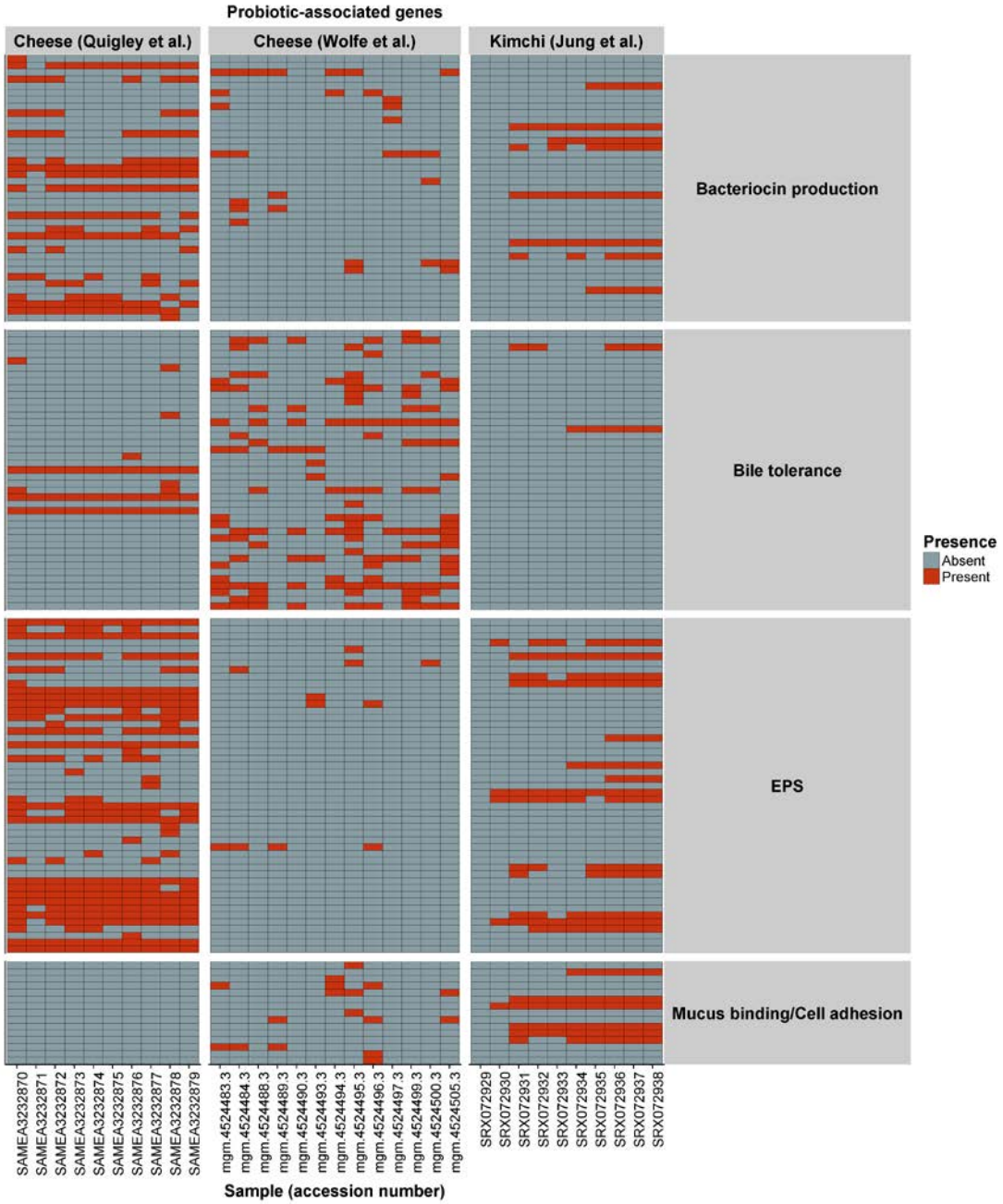


Figure 3: Binary heatmap showing the presence/absence of genes associated with probiotic action in cheese and kimchi metagenomes, as determined by HUMAnN2.

mucus binding proteins, bile salt hydrolases, bile salt symporters, and bacteriocins/prebacteriocins were widespread in the 14 cheese varieties investigated (Figure 3). In addition, we observed several instances where multiple genes were assigned to individual species (Table S2). We identified similar genes in kimchi (Figure 3), although HUMAnN2 was unable to assign them to individual species because of the lower sequencing depth of those samples.

Volatile profiling and sensory analysis of kefir milk

GCMS was used to determine the volatile profile of kefir milk after 0, 8 and 24 hours of fermentation. 39 volatile compounds that could contribute to flavour were identified and semi-quantified in kefir milks produced with each of the three kefir grains. These consisted of 9 ketones, 7 aldehydes, 6 esters, 8 alcohols, 5 carboxylic acids and 2 sulphur compounds (Table 1). The results of the volatile analysis are presented in Figure 4. The levels of all of the detected compounds increased after 0 hours, apart from 1-pentanol, pentanal, hexanal, heptanal, heptanol, acetone and 2-butanone (Figure 4).

Sensory acceptance evaluation and ranking descriptive analysis (RDA) were performed on the Fr1 and Ick kefir milks after 24 hour fermentations. These revealed perceptible differences between the milks. Specifically, Fr1 samples had a more likeable, buttery flavour whereas Ick samples had a less likeable but fruity flavour (Figure S5). These results confirm that the volatile profile data is consistent with subsequent flavour.

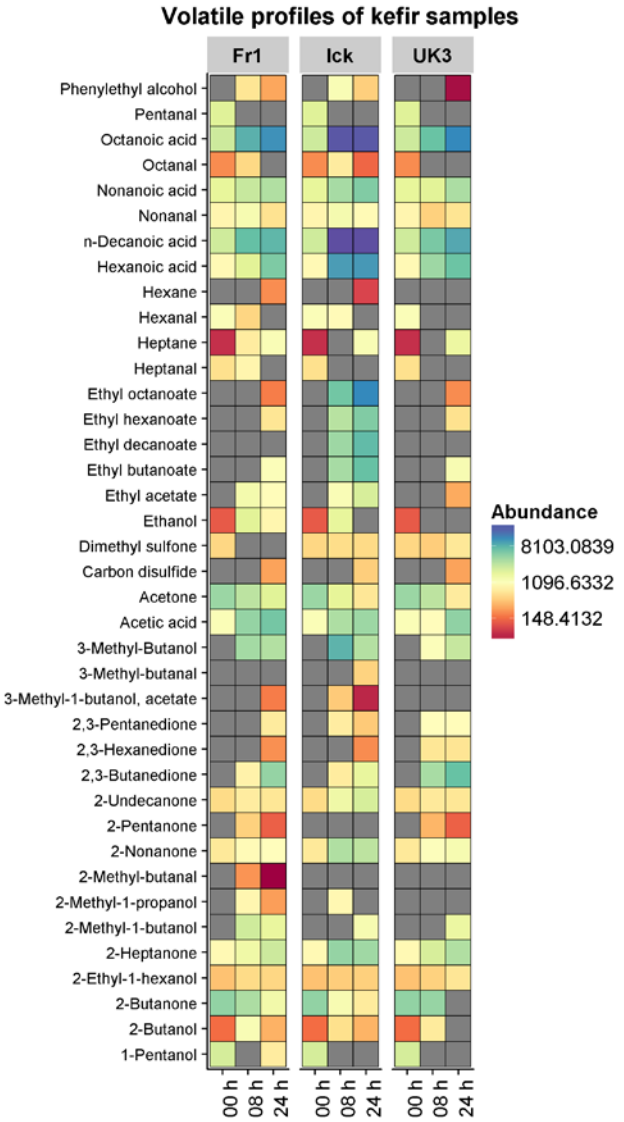


Figure 4: Facetted heatmap showing the volatile profiles of the kefirs Fr1, Ick and UK3 at 0 h, 8 h and 24 h of fermentation.

Correlations between microbial taxa and volatile compounds

The Spearman rank correlation test was used to identify correlations between the levels of individual taxa and flavour compounds. At the genus level, based on amplicon sequencing results, there were strong correlations between *Lactobacillus* and carboxylic acids, esters and 3-methyl-1-butanol; between *Saccharomyces* and carboxylic acids, and esters; between *Acetobacter* and acetic acid, 2-methyl-1-butanol, and 2,3-butanedione; between *Leuconostoc* and 2,3-butanedione; and between *Kazachstania* and acetic acid, 2-methyl-1-butanol, 2,3-butanedione, 2,3-pentanedione, and 2,3-hexanedione (Table S3). At the bacterial species level, based on results from Kraken, there were strong correlations between *Lb. kefiranofaciens* and carboxylic acids and esters; between *A. pasteurianus* and carboxylic acids, and 2,3-butanedione; and between *L. mesenteroides* and 2,3-butanedione. At the fungal species level, based on results from MetaPhlan2, there were strong correlations between *S. cerevisiae* and alcohols and esters (Table 3, Figure 5). In summary, correlations were found between compounds associated with vinegary-flavours and *A. pasteurianus*, cheesy-flavours and *Lb. kefiranofaciens*, buttery-flavours and *L. mesenteroides*, and fruity-flavours with *Lb. kefiranofaciens* and *S. cerevisiae.*

Impact of supplementing kefir with kefir isolates

The consequences of adding *Lb. kefiranofaciens* NCFB 2797 to Fr1 was investigated, since this kefir had a low indigenous *Lb. kefiranofaciens* population. GCMS revealed that this addition caused increases in the levels of the esters ethenyl acetate (by 59.15%), ethyl acetate (100%), methyl-3-butyrate (26.83%), and 2-methylbutyl-acetate (11.44%), and the ketone 2-heptanone (65.86%). In contrast, the

Table 3: Summary of strong positive correlations (R>0.5) identified between the relative abundance of species and the level of metabolites in kefir.

Species	Compound(s)	R values	Unadjusted p-value	FDR adjusted p-value
Leuconostoc mesenteroides	2,3-Butanedione	0.79	0.0005	0.011
Lactobacillus kefiranofaciens	2-Nonanone	0.79	0.0005	0.011
Lactobacillus helveticus	Acetic acid	0.75	0.0013	0.017
Leuconostoc mesenteroides	2-Methyl-1-butanol	0.74	0.0015	0.017
Lactobacillus kefiranofaciens	Hexanoic acid	0.71	0.0033	0.024
Lactobacillus kefiranofaciens	2-Heptanone	0.71	0.0033	0.024
Lactobacillus kefiranofaciens	Octanoic acid	0.7	0.004	0.024
Lactobacillus kefiranofaciens	Acids	0.68	0.0049	0.024
Lactobacillus.kefiranofaciens	n-Decanoic acid	0.68	0.0049	0.024
Saccharomyces cerevisiae	Nonanal	0.66	0.008	0.035
Lactobacillus kefiranofaciens	Ethyl decanoate	0.65	0.0089	0.035
Lactobacillus kefiranofaciens	Esters	0.64	0.0099	0.035
Saccharomyces cerevisiae	Ethyl acetate	0.63	0.011	0.035
Saccharomyces cerevisiae	3-Methyl-Butanol	0.63	0.0118	0.035
Acetobacter pasteurianus	2-Methyl-1-butanol	0.63	0.0126	0.035
Saccharomyces cerevisiae	Phenylethyl alcohol	0.62	0.0134	0.035
Saccharomyces cerevisiae	Octanal	0.62	0.0141	0.035
Acetobacter pasteurianus	Nonanoic acid	0.6	0.0169	0.04
Lactobacillus kefiranofaciens	Phenylethyl alcohol	0.6	0.0179	0.04
Saccharomyces cerevisiae	Alcohols	0.59	0.0203	0.04
Acetobacter pasteurianus	Acetic acid	0.59	0.0206	0.04
Lactobacillus helveticus	2,3-Butanedione	0.57	0.0255	0.04
Acetobacter pasteurianus	Ethyl butanoate	0.57	0.0268	0.04
Lactobacillus kefiranofaciens	Ethyl hexanoate	0.57	0.0279	0.04
Lactobacillus.helveticus	3-Methyl-Butanol	0.56	0.0283	0.04
Saccharomyces cerevisiae	Ethyl decanoate	0.56	0.0283	0.04
Lactobacillus kefiranofaciens	Ethyl butanoate	0.56	0.0292	0.04
Acetobacter pasteurianus	Ethyl hexanoate	0.56	0.0314	0.04
Lactobacillus kefiranofaciens	Nonanoic acid	0.56	0.0315	0.04
Lactobacillus kefiranofaciens	2-Undecanone	0.56	0.0315	0.04
Lactobacillus kefiranofaciens	Ethyl octanoate	0.55	0.0321	0.04
Lactobacillus kefiranofaciens	3-Methyl-Butanol	0.55	0.0321	0.04
Acetobacter pasteurianus	Acids	0.55	0.0324	0.04
Acetobacter pasteurianus	Hexanoic acid	0.54	0.0368	0.044
Acetobacter pasteurianus	2,3-Butanedione	0.54	0.0375	0.044
Acetobacter pasteurianus	Ethyl acetate	0.54	0.0381	0.044
Saccharomyces cerevisiae	Esters	0.53	0.0413	0.046
Leuconostoc mesenteroides	Acetic acid	0.53	0.0419	0.046
Lactobacillus helveticus	2-Methyl-1-butanol	0.53	0.0433	0.046
Saccharomyces cerevisiae	2-Undecanone	0.53	0.0435	0.046
Lactobacillus kefiranofaciens	Ethyl acetate	0.52	0.0478	0.048
Acetobacter pasteurianus	Octanoic acid	0.52	0.0478	0.048

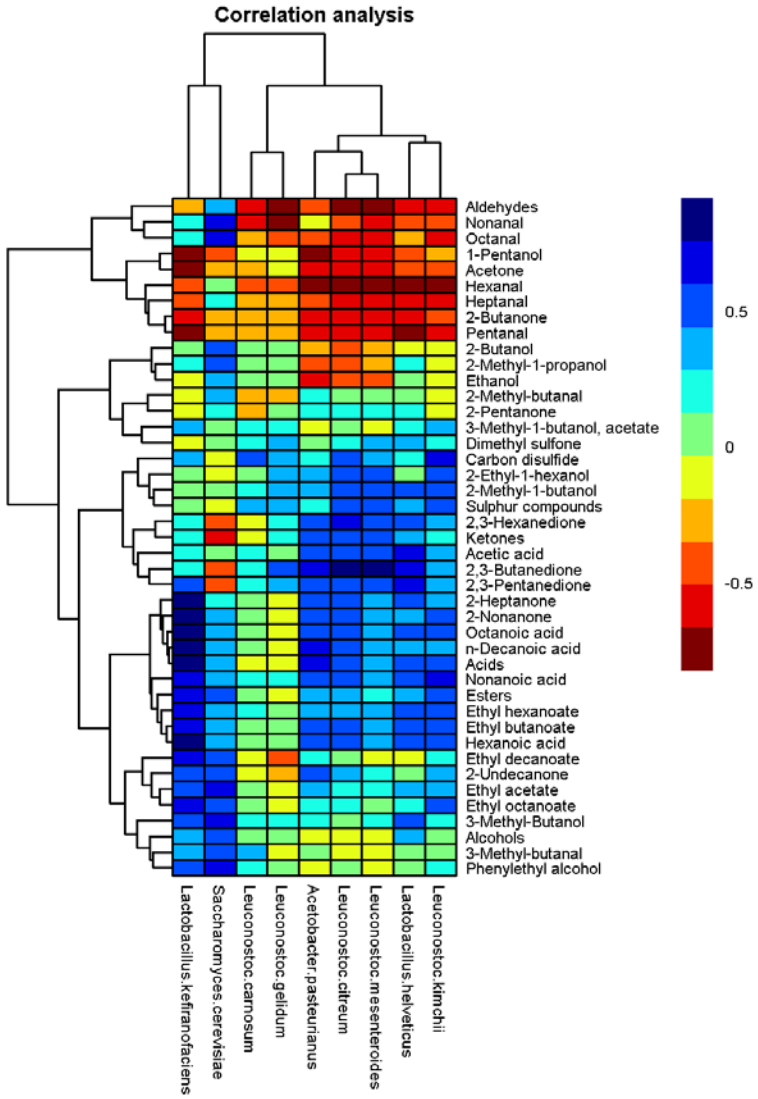


Figure 5: Hierarchically clustered heatmap showing correlations between the relative abundances of microbial species and the levels of volatile compounds in kefir samples. The colour of each tile of the heatmap indicates the type/strength of the correlation for a given species/compound combination, as indicated by the colour key.

addition of *L. mesenteroides* DPC 7047 to Ick, a kefir with a low indigenous *L. mesenteroides* population, resulted in increases in the levels of acetic acid (168.28%) and 2,3-butanediol (14.91%), a precursor to 2,3-butanedione (Table S4). Despite changes in volatile profile, there were no perceptible changes in flavour (Figure S5).

DISCUSSION

Many traditional fermented foods have been reported to have health benefits (44, 45). These foods are often produced on a small-scale, artisanal basis. However, the increased demand for health-promoting foods among the public presents an opportunity to bring traditional fermented foods to a wider audience and serves as an incentive to optimise starter cultures for the mass production of fermented foods with enhanced sensory qualities (46). In recent years, genetic characterisation has been increasingly employed to guide starter culture development for numerous fermented foods, including wines, beers, cocoa, and meats (47-50). Similarly, integrated molecular 'omics' approaches (51) have emerged as powerful methods of investigating the microbial dynamics of food fermentations with the aim of optimising processes like flavour production (52). In this study, we combined compositional and shotgun

DNA sequencing with GCMS and flavour analysis to predict microbes involved in the production of different flavour compounds in kefir.

We identified significant correlations between the abundances of particular microbial genera and species and the levels of different volatile compounds, and showed that the microbes in kefir had genes necessary for the production of these compounds. Specifically, *Acetobacter pasteurianus* correlated with acetic acid which is associated with vinegary-flavours; *Lb. kefiranofaciens* correlated with carboxylic acids and ketones associated with cheesy-flavours, and esters associated with fruity-flavours; *L. mesenteroides* correlated with 2,3-butanedione, which is associated with buttery-flavours, and acetic acid; and *S. cerevisiae* correlated with esters. Sensory analysis revealed that Fr1, a kefir high in *L. mesenteroides*, had a likeable buttery flavour, whereas Ick, a kefir high in *Lb. kefiranofaciens*, had a less likeable but fruity flavour. Thus, our data suggested a causal relationship between specific taxa and flavour characteristics, which was subsequently supported by experimentally manipulating the kefir community. In line with predictions, adding *Lb. kefiranofaciens* NCFB 2797 to Fr1 resulted in increases in the levels of 2-heptanone and esters, whereas the addition of *L. mesenteroides* DPC 7047 to Ick resulted in increases in the levels of acetic acid and 2,3-butanediol, a precursor to 2,3-butanedione. However,

sensory analysis indicated that these changes were imperceptible, and therefore higher inoculum levels might be necessary to change flavour.

Based on these results, we predict that the final flavour of kefir can be manipulated by altering the ratio of microbes in the grain. Unfortunately, to date, it has not been possible to artificially reconstruct kefir grains in the laboratory and this might hamper the practical application of our findings. However, we propose that the approach outlined here can be used to accelerate the development of superior multi-strain starter cultures to improve the flavour of a variety of fermented foods.

From a systems biology perspective, our work confirms that kefir is suitable as a model microbial community. There are two advantages to using the kefir model, rather than other fermented foods, in this way. Firstly, kefir contains fewer species, and so is a simpler environment in which to investigate how microbial communities are formed. Secondly, kefir is quick and easy to produce; with the fermentation taking just 24 hours when incubated at room temperature. In addition, others have demonstrated that kefir is a highly culturable system and, indeed, all of the species that were detected at a relative abundance >1% at 8 and 24 hours across the examined kefirs have been isolated previously (53).

Ultimately, Kraken and MetaPhlAn2 showed that the microbial population of kefir was dominated by *Lb. kefiranofaciens* at 8 hours of

fermentation. However, between 8 and 24 hours, there was a fall in the relative abundance of *Lb. kefiranofaciens* and *L. mesenteroides* superseded it as the dominant species. The shift from *Lb. kefiranofaciens* to *L. mesenteroides* is similar to patterns of microbial succession seen in other fermented foods (54, 55). We propose that kefir could be a particularly appropriate model community in which to determine the driving-forces behind microbial succession. Early colonising bacteria in other fermentations have been reported to modify the environment in such a way as to make it more suitable for the growth of other bacteria, thus driving succession (5), and this could explain the observed shift that occurs during kefir fermentation. Our HUMAnN2 results revealed that genes involved in aromatic amino acid biosynthesis were assigned to *L. mesenteroides* but not *Lb. kefiranofaciens.* This may be significant because free amino acid analysis showed that there was a significant decrease in the levels of tyrosine in kefir between 8 and 24 hours (Supplemental Results). It is possible that its ability to synthesise tyrosine underlies the increased prevalence of *L. mesenteroides,* relative to *Lb. kefiranofaciens,* in the latter stages of fermentation. Future work will focus on investigating the effect of modifying the levels of tyrosine on the microbiota and volatile profile of kefir. Thus, a 'kefir model' has the potential to yield insights into the effects of nutrient availability on

microbial succession and metabolite production in other, more complicated, environments.

Finally, we showed that *Lb. kefiranofaciens* has genes which encode proteins that are considered to be important for probiotic action, including exopolysaccharide synthesis proteins, bile salt transporters, mucus binding proteins and bacteriolysins (56, 57). The presence of these genes suggests that the *Lb. kefiranofaciens* strains present in these kefirs have the potential to survive gastric transit, colonise the gut and inhibit the growth of pathogens. Indeed, previous studies using mice have shown that *Lb. kefiranofaciens* protects against enterohemorrhagic *Escherichia coli* infection (58). Further analysis of shotgun metagenomic data from cheese and kimchi indicated that similar genes are present in other fermented foods. Our findings are consistent with previous observations that fermented food-borne microbes can colonise the gut (59), and support designating some fermented foods, like kimchi, as ‘probiotic foods’ (45).

In summary, in this study it has been demonstrated that a combined metagenomics and metabolomics approach can potentially be used to identify the microbes from a particular environment that are responsible for the production of certain metabolites, using the production of flavour compounds during kefir fermentation as a model. Furthermore, we have

provided additional evidence of the use of microbial fermentations to provide valuable insights into the dynamics of microbial succession and, in the process, identified genes in *Lb. kefiranofaciens* that potentially confer important probiotic traits. To conclude, our analyses confirm the value of using kefir as a model microbial community, while also providing a valuable insight into the microbiology of this natural health promoting beverage.

References

1. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG.** 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences **109:**21390-21395.
2. **Consortium HMP.** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486:**207.
3. **Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences **103:**12115-12120.
4. **Faust K, Raes J.** 2012. Microbial interactions: from networks to models. Nature Reviews Microbiology **10:**538-550.
5. **Wolfe BE, Dutton RJ.** 2015. Fermented Foods as Experimentally Tractable Microbial Ecosystems. Cell **161:**49-55.
6. **Wolfe BE, Button JE, Santarelli M, Dutton RJ.** 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell **158:**422-433.
7. **Nielsen B, Gürakan GC, Ünlü G.** 2014. Kefir: a multifaceted fermented dairy product. Probiotics and Antimicrobial Proteins **6:**123-135.
8. **Güzel-Seydim Z, Seydim A, Greene A, Bodine A.** 2000. Determination of organic acids and volatile flavor substances in kefir during fermentation. Journal of Food composition and Analysis **13:**35-43.

9. **Marsh AJ, O'Sullivan O, Hill C, Ross RP, Cotter PD.** 2013. Sequencing-based analysis of the bacterial and fungal composition of kefir grains and milks from multiple sources. PloS One **8:**e69171.

10. **Nalbantoglu U, Cakar A, Dogan H, Abaci N, Ustek D, Sayood K, Can H.** 2014. Metagenomic analysis of the microbial community in kefir grains. Food Microbiology **41:**42-51.

11. **Garofalo C, Osimani A, Milanović V, Aquilanti L, De Filippis F, Stellato G, Di Mauro S, Turchetti B, Buzzini P, Ercolini D.** 2015. Bacteria and yeast microbiota in milk kefir grains from different Italian regions. Food Microbiology **49:**123-133.

12. **Korsak N, Taminiau B, Leclercq M, Nezer C, Crevecoeur S, Ferauche C, Detry E, Delcenserie V, Daube G.** 2015. Short communication: evaluation of the microbiota of kefir samples using metagenetic analysis targeting the 16S and 26S ribosomal DNA fragments. Journal of Dairy Science **98:**3684-3689.

13. **de LeBlanc AdM, Matar C, Farnworth E, Perdigon G.** 2007. Study of immune cells involved in the antitumor effect of kefir in a murine breast cancer model. Journal of Dairy Science **90:**1920-1928.

14. **Lee M-Y, Ahn K-S, Kwon O-K, Kim M-J, Kim M-K, Lee I-Y, Oh S-R, Lee H-K.** 2007. Anti-inflammatory and anti-allergic effects of kefir in a mouse asthma model. Immunobiology **212:**647-654.

15. **Rodrigues KL, Caputo LRG, Carvalho JCT, Evangelista J, Schneedorf JM.** 2005. Antimicrobial and healing activity of kefir and kefiran extract. International Journal of Antimicrobial Agents **25:**404-408.

16. **Hertzler SR, Clancy SM.** 2003. Kefir improves lactose digestion and tolerance in adults with lactose maldigestion. Journal of the American Dietetic Association **103:**582-587.

17. **Liu J-R, Wang S-Y, Chen M-J, Chen H-L, Yueh P-Y, Lin C-W.** 2006. Hypocholesterolaemic effects of milk-kefir and soyamilk-kefir in cholesterol-fed hamsters. British Journal of Nutrition **95:**939-946.

18. **Bolla P, Abraham A, Pérez P, de los Angeles Serradell M.** 2015. Kefir-isolated bacteria and yeasts inhibit Shigella flexneri invasion and modulate pro-inflammatory response on intestinal epithelial cells. Beneficial Microbes**:**1-8.

19. **Leite AMO, Miguel MAL, Peixoto RS, Ruas-Madiedo P, Paschoalin VMF, Mayo B, Delgado S.** 2015. Probiotic potential of selected lactic acid bacteria strains isolated from Brazilian kefir grains. Journal of Dairy Science **98:**3622-3632.

20. **Carasi P, Racedo S, Jacquot C, Romanin D, Serradell M, Urdaci M.** 2015. Impact of kefir derived Lactobacillus kefiri on the mucosal immune response and gut microbiota. Journal of Immunology Research **2015**.

21. **De Montijo-Prieto S, Moreno E, Bergillos-Meca T, Lasserrot A, Ruiz-López M-D, Ruiz-Bravo A, Jiménez-Valera M.** 2015. A Lactobacillus plantarum strain isolated from kefir protects against intestinal infection with Yersinia enterocolitica O9 and modulates immunity in mice. Research in Microbiology **166:**626-632.

22. **Bolla PA, Carasi P, Bolla MdlA, De Antoni GL, Serradell MdlA.** 2013. Protective effect of a mixture of kefir-isolated lactic acid bacteria and yeasts in a hamster model of Clostridium difficile infection. Anaerobe **21:**28-33.

23. **Fouhy F, Guinane CM, Hussey S, Wall R, Ryan CA, Dempsey EM, Murphy B, Ross RP, Fitzgerald GF, Stanton C.** 2012. High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. Antimicrobial Agents and Chemotherapy **56:**5811-5820.

24. **Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, Claesson MJ.** 2016. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PLoS One **11:**e0148028.

25. **Orgiazzi A, Lumini E, Nilsson RH, Girlanda M, Vizzini A, Bonfante P, Bianciotto V.** 2012. Unravelling soil fungal communities from different Mediterranean land-use backgrounds. PloS One **7:**e34847.

26. **Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Ross RP, O'Callaghan G, Plant BJ, Stanton C.** 2015. The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. PloS One **10:**e0119355.

27. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7:**335-336.

28. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.

29. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of Molecular Biology **215:**403-410.

30. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research **41:**D590-D596.

31. **Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G.** 2012. Reference databases for taxonomic assignment in metagenomics. Briefings in Bioinformatics**:**bbs036.

32. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

33. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methe B, Schloss PD, Gevers D, Mitreva M, Huttenhower C.** 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol **8:**e1002358.

34. **Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH.** 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23:**1282-1288.

35. **Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C.** 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods **9:**811-814.

36. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15:**R46.

37. **Team RC.** 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.

38. **Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C.** 2011. Metagenomic biomarker discovery and explanation. Genome Biology **12:**1.

39. **Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N.** 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ **3:**e1029.

40. **Quigley L, McCarthy R, O'Sullivan O, Beresford TP, Fitzgerald GF, Ross RP, Stanton C, Cotter PD.** 2013. The microbial content of raw and pasteurized cow milk as determined by molecular approaches. Journal of Dairy Science **96:**4928-4937.

41. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

42. **Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A.** 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research **42:**D459-D471.

43. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT.** 2000. Gene Ontology: tool for the unification of biology. Nature Genetics **25:**25-29.

44. **Marsh AJ, Hill C, Ross RP, Cotter PD.** 2014. Fermented beverages with health-promoting potential: past and future perspectives. Trends in Food Science & Technology **38:**113-124.

45. **Park K-Y, Jeong J-K, Lee Y-E, Daily III JW.** 2014. Health benefits of kimchi (Korean fermented vegetables) as a probiotic food. Journal of Medicinal Food **17:**6-20.

46. **Bachmann H, Pronk JT, Kleerebezem M, Teusink B.** 2015. Evolutionary engineering to enhance starter culture performance in food fermentations. Current Opinion in Biotechnology **32:**1-7.

47. **Bellon JR, Yang F, Day MP, Inglis DL, Chambers PJ.** 2015. Designing and creating Saccharomyces interspecific hybrids for improved, industry relevant, phenotypes. Applied Microbiology and Biotechnology **99:**8597-8609.

48. **Steensels J, Snoek T, Meersman E, Nicolino MP, Voordeckers K, Verstrepen KJ.** 2014. Improving industrial yeast strains: exploiting natural and artificial diversity. FEMS Microbiology Reviews **38:**947-995.

49. **Meersman E, Steensels J, Paulus T, Struyf N, Saels V, Mathawan M, Koffi J, Vrancken G, Verstrepen KJ.** 2015. Breeding strategy to generate robust yeast starter cultures for cocoa pulp fermentations. Applied and Environmental Microbiology **81:**6166-6176.

50. **Flores M, Corral S, Cano-García L, Salvador A, Belloch C.** 2015. Yeast strains as potential aroma enhancers in dry fermented sausages. International Journal of Food Microbiology **212:**16-24.

51. **Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C.** 2015. Sequencing and beyond: integrating

molecular'omics' for microbial community profiling. Nature Reviews Microbiology **13:**360-372.

52. **Wang Z-M, Lu Z-M, Shi J-S, Xu Z-H.** 2016. Exploring flavour-producing core microbiota in multispecies solid-state fermentation of traditional Chinese vinegar. Scientific Reports **6:**26818.

53. **Bourrie BCT, Willing BP, Cotter PD.** 2016. The Microbiota and health promoting characteristics of the fermented beverage Kefir. Frontiers in Microbiology **7**.

54. **Jeong SH, Jung JY, Lee SH, Jin HM, Jeon CO.** 2013. Microbial succession and metabolite changes during fermentation of dongchimi, traditional Korean watery kimchi. International Journal of Food Microbiology **164:**46-53.

55. **Jung JY, Lee SH, Lee HJ, Jeon CO.** 2013. Microbial succession and metabolite changes during fermentation of saeu-jeot: traditional Korean salted seafood. Food Microbiology **34:**360-368.

56. **Bermudez-Brito M, Plaza-Díaz J, Muñoz-Quezada S, Gómez-Llorente C, Gil A.** 2012. Probiotic mechanisms of action. Annals of Nutrition and Metabolism **61:**160-174.

57. **Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, O'Toole PW.** 2009. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nature Reviews Microbiology **7:**61-71.

58. **Chen Y, Lee T, Hong W, Hsieh H, Chen M.** 2013. Effects of Lactobacillus kefiranofaciens M1 isolated from kefir grains on enterohemorrhagic Escherichia coli infection using mouse and intestinal cell models. Journal of Dairy Science **96:**7467-7477.

59. **David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA.** 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature **505:**559-563.

SUPPLEMENTAL MATERIAL

Supplemental materials and methods

Volatile profiling of spiked kefir by GCMS

2 g of sample was added to 20 ml screw capped SPME vial and equilibrated to 40 °C for 10 mins with pulsed agitation of 5 sec at 500 rpm. The samples were analysed in triplicate. Sample introduction was accomplished using a CTC Analytics CombiPalAutosampler.

A single 50/30 µm Carboxen^{TM}/divinylbenzene/polydimethylsiloxane (DVB/CAR/PDMS) fiber was used. The SPME fiber was exposed to the headspace above the samples for 20 min at depth of 1 cm at 40°C. The fiber was retracted and injected into the GC inlet and desorbed for 2 min at 250°C. Injections were made on an Shimadzu 2010 Plus GC with an Agilent DB-5 (60 m x 0.25 mm x 0.25 µm) column using a split/splitless injector in splitless mode with a merlin microseal. The temperature of the column oven was set at 35°C, held for 0.5 min, increased at 6.5°C/min to 230°C then increased at 15°C/min to 320°C, yielding at total GC run time of 41.5 min. The carrier gas was helium held at a constant pressure of 23 psi. The detector was a Shimadzu TQ8030 mass spectrometer detector, ran in single quad mode. The ion source temperature was 220°C and the interface temperature were set at 280°C and the MS mode was electronic ionization (-70 v) with the mass range scanned between 35 and 250 amu. Compounds were identified using mass spectra comparisons to the NIST 2014 mass spectral library and in-house library created in Chem Solutions software (Shimadzu, Japan) with target and qualifier ions and linear retention indices for each compound. Final data processing was undertaken using TargetView deconvolution software (Markes International Ltd,

UK). An auto-tune of the GCMS was carried out prior to the analysis to ensure optimal GCMS performance. A set of external standards was ran at the start and end of the sample set and abundances were compared to known amounts to ensure that both the SPME extraction and MS detection was performing within specification.

Sensory acceptance evaluation of spiked and non-spiked kefir milks

Twenty five naïve assessors were recruited in University College Cork, Ireland for sensory acceptance evaluation of spiked and non-spiked kefir milks. Age range of assessors was 21-48 years old. The selection criteria for assessors were availability and motivation to participate on all days of the experiment. Assessors used sensory Hedonic descriptors (Table S5) on 11 different kefir samples. Sensory analysis was carried out in panel booths conforming to international standards (ISO 8589:2007). All samples were stored at -20˚C until required. Samples were then held at refrigeration temperatures overnight (4ºC), before monadic presentation to the consumer panel at ambient temperatures (21°C) and coded with a randomly selected 3 digit code. A maximum of six samples were presented at each session. Each assessor was provided with deionised water and instructed to cleanse their palates between tastings asked to assess the attributes, according to a ten-point scale. The order of the presentation of all test samples was randomized to prevent first order and carryover effects.

Ranking descriptive analysis (RDA) of spiked and non-spiked kefir milks

Ten panellists were recruited in University College Cork, Ireland. Age range of assessors was 25-45 years old. Selection criteria for panellists were availability and motivation to participate on all days of the experiment and that they were familiar with kefir as a product. All panellists had participated in dairy descriptive profiles in

the past and were well versed in the sensory experimental protocol. Panellists were trained using the sensory Intensity descriptors (Table S5). Ranking Descriptive analysis (RDA) [60, 61] was carried out in panel booths conforming to international standards (ISO 8589:2007) on the 11 Kefir samples to be tested. All samples were stored at -20˚C until required. Samples were then held at refrigeration temperatures overnight (4ºC), before presentation to the panel at ambient temperatures (21°C) and coded with a randomly selected 3 digit code. The Kefir samples were immediately served to panellists simultaneously in separate sessions for Fr1 and Ick variants. Each assessor was provided with deionised water and instructed to cleanse their palates between tastings. Additionally, each assessor was presented with samples and asked to rank the intensity of the attributes, according a 10 cm line scale ranging from 0 (none) at the left to 10 (extreme) at the right and rating subsequently scored in cm from left (Table S5). The order of the presentation of all test samples was randomized to prevent first order and carryover effects.

Statistical analysis of sensory analysis data

For evaluating the results of the RDA and the sensory acceptance test, ANOVA-Partial Least Squares regression (APLSR) was used to process the data accumulated using Unscrambler software version10.3. The X-matrix was designed as 0/1 variables for sample and the Y-matrix sensory variables.

Free amino acid analysis

The aromatic amino acids phenylalanine and tyrosine were quantified in the milk samples using the method described by McDermott *et al.* [62].

Supplemental results

Sequencing results

16S rRNA gene sequencing generated 5,545,825 reads in total and an average of 210,136 reads per sample, while ITS gene sequencing generated 3,498,902 reads in total and an average of 291,567 reads per sample. Whole metagenome sequencing generated a total of 22,983,010 reads and an average of 1,209,632 reads per sample.

Free amino acid analysis results

Free amino acid analysis showed that the levels of phenylalanine increased from 0.63 to 0.98 nmol/ml between 8 and 24 hours, but the Wilcoxon signed rank test indicated that this increase was not statistically significant ($p=0.064$). In contrast, the levels of tyrosine decreased from 4.18 to 1.42 nmol/ml between 8 and 24 hours, and the Wilcoxon signed rank test indicated that this decrease was statistically significant ($p=0.018$).

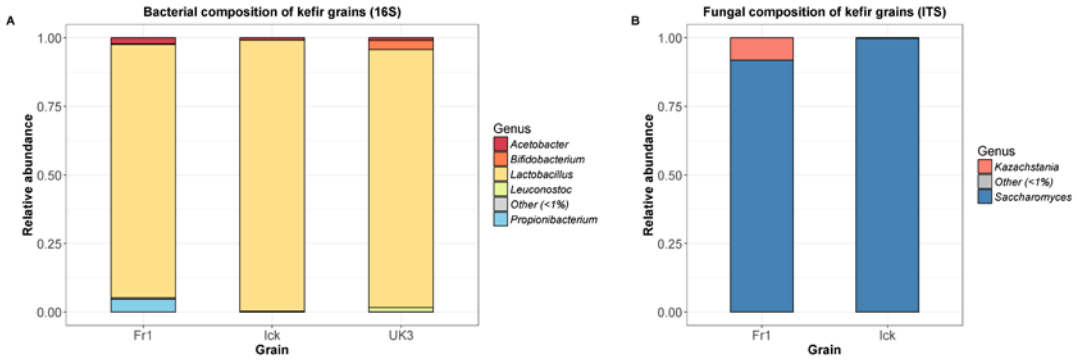


Figure S1: The (a) bacterial and (b) fungal composition of kefir grains, as determined by amplicon sequencing. Note that we were unable to generate an ITS amplicon for the UK3 sample.

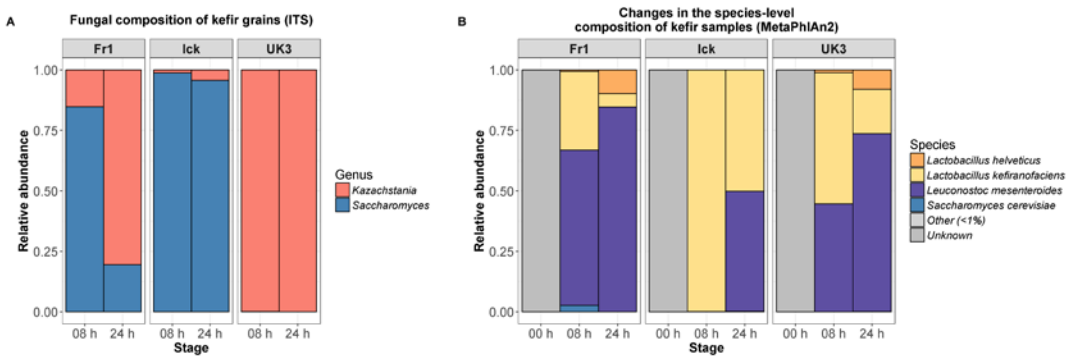


Figure S2: Stacked bar charts presenting (a) the fungal composition of kefir samples after 0, 8 and 24 hours of fermentation, as determined by ITS gene sequencing, and (b) the microbial composition of kefir samples after 0, 8 and 24 hours of fermentation, as determined by MetaPhlAn2.

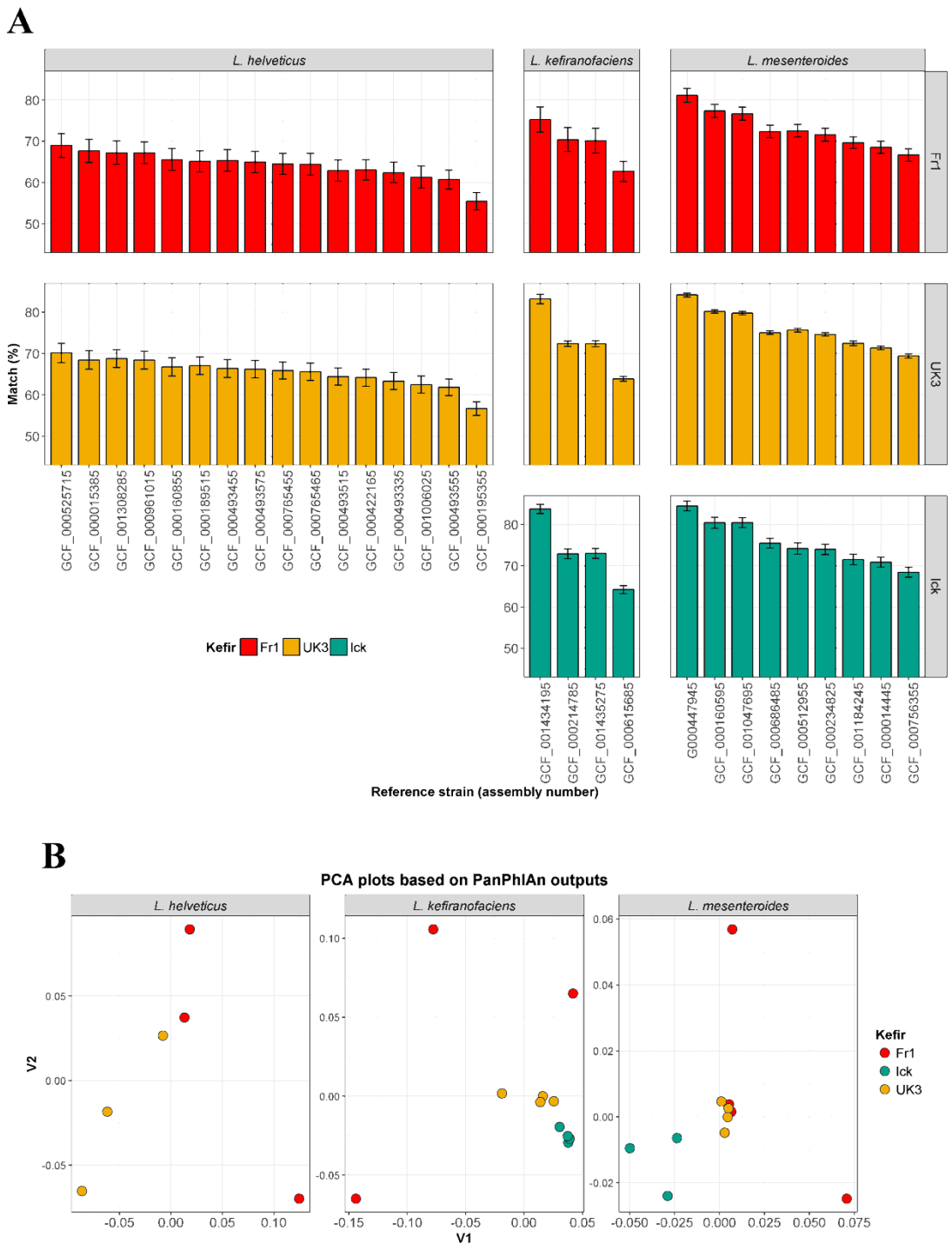


Figure S3: PanPhlAn analysis of the dominant bacterial species detected in kefir. (A) Bar plots displaying the percentage of pangenome gene families shared between the detected strains and their respective reference genomes. (B) Principal-component analysis (PCA) plot based on the presence/absence of pangenome gene families in detected strains.

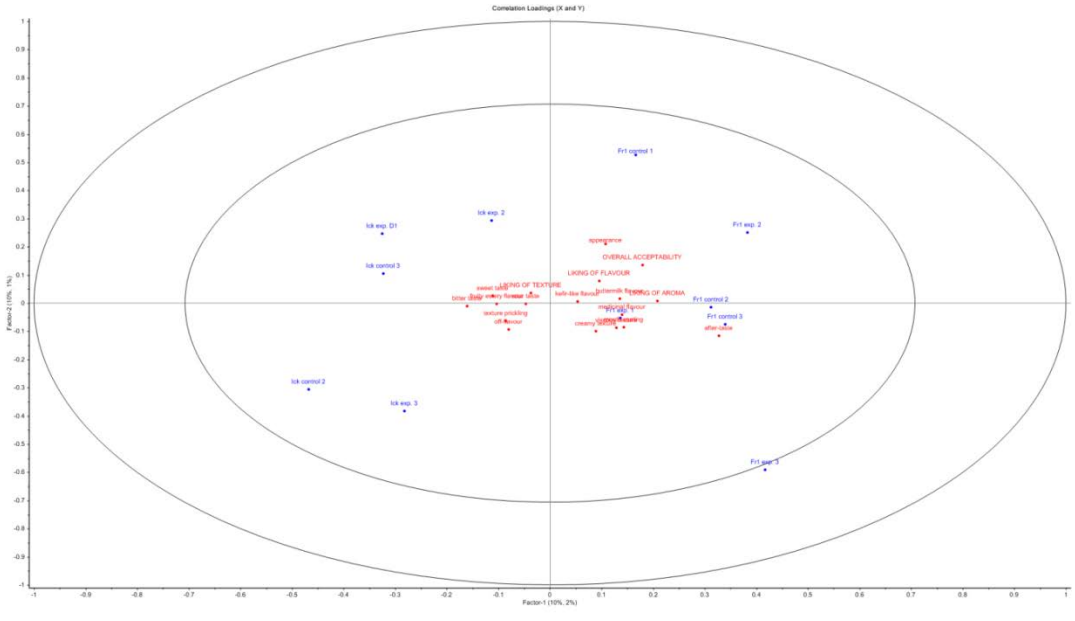


Figure S4: ANOVA-Partial Least Squares Regression (ASLPR, PCs 1-2) plot for spiked and non-spiked kefir samples presenting Sensory Acceptance and Ranking Descriptive Analysis data.

Table S1: Absolute abundances of bacteria and fungi in kefir samples after 0, 8 and 24 hours of fermentation, as determined by quantitative PCR (qPCR) measurements.

Sample	Total fungi (ng of fungal DNA)	Total bacteria (copies of 16S rRNA gene)
Milk 0 h	0.0016	1.78E+05
Fr1 08 h	0.1386	7.64E+07
Fr1 24 h	0.2179	2.49E+08
Ick 08 h	0.0542	2.62E+08
Ick 24 h	0.0972	1.63E+08
UK3 08 h	0.0896	7.86E+07
Uk3 24 h	0.3933	4.22E+08

Table S2: Microbial species from cheese samples that contain two or more genes associated with probiotic action, as determined by HUMAnN2.

Species	UniRef50 genes
Lactobacillus casei paracasei	UniRef50_K0N6Y4: Exopolysaccharide biosynthesis protein
	UniRef50_B3WA97: Prebacteriocin
	UniRef50_D4MCD9: Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase
	UniRef50_S6CA35: Truncated bacteriocin ABC transporter ATP-binding and permease components
	UniRef50_F0TGY1: Exopolysaccharide biosynthesis protein
	UniRef50_K6PVY4: Prebacteriocin
	UniRef50_S6BTB5: Exopolysaccharide synthesis protein
	UniRef50_S2U8B2: Exopolysaccharide phosphogalactosyltransferase
Streptococcus thermophilus	UniRef50_S6C1D3: Truncated bacteriocin ABC transporter ATP-binding and permease components
	UniRef50_T0T7J2: Pleiotropic regulator of exopolysaccharide synthesis, competence and biofilm formation Ftr, nREfamily
	UniRef50_F8DFF3: Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase
	UniRef50_Q03K73: Exopolysaccharide biosynthesis protein related to N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase
	UniRef50_W4KQU9: Pore-forming peptide bacteriocin
	UniRef50_W7VD14: Exopolysaccharide biosynthesis protein
	UniRef50_UPI000046DD7F: exopolysaccharide biosynthesis protein, truncated, partial
	UniRef50_R5ZYR0: Serine (Threonine) dehydratase involved in lantibiotic biosynthesis
	UniRef50_F5X0K2: Glycosyltransferase in exopolysaccharide biosynthesis
	UniRef50_Q5LZN9: Exopolysaccharide biosynthesis protein, sugar transferase
	UniRef50_Q5LZP0: Exopolysaccharide polymerization protein
	UniRef50_Q5LZP3: Exopolysaccharide biosynthesis protein
Lactobacillus delbrueckii	UniRef50_F0K085: Exopolysaccharide biosynthesis protein
	UniRef50_W5XGS2: Exopolysaccharide biosynthesis protein
	UniRef50_F6CC46: Exopolysaccharide biosynthesis protein
	UniRef50_D5GZ53: Response regulator bacteriocinproduction-related
	UniRef50_F0TGY1: Exopolysaccharide biosynthesis protein
	UniRef50_F6CC46: Exopolysaccharide biosynthesis protein
	UniRef50_W5XGS2: Exopolysaccharide biosynthesis protein
Clostridium tyrobutyricum	UniRef50_G7M1R9: Capsular exopolysaccharide family
	UniRef50_R4K4E3: Exopolysaccharide biosynthesis protein
	UniRef50_W6NKH4: Linocin_M18 bacteriocin protein
Lactococcus raffinolactis	UniRef50_S6CA35: Truncated bacteriocin ABC transporter ATP-binding and permease components
	UniRef50_I7JFK8: Exopolysaccharide biosynthesis protein
	UniRef50_I7LPF2: Exopolysaccharide biosynthesis protein
Brevibacterium linens	UniRef50_UPI0001BC2DFF: bacteriocin ABC transporter ATP-binding protein
	UniRef50_D6ZIG8: Bile acid transporter
	UniRef50_K9B0K5: Collagen adhesion protein
	UniRef50_A0JWH3: L-carnitine dehydratase/bile acid-inducible protein F
	UniRef50_UPI000050FBF9: L-carnitine dehydratase/bile acid-inducible protein F
	UniRef50_A0A022L0W5: Bile acid:sodium symporter
Enterococcus faecalis	UniRef50_R4A735: Collagen adhesion protein
	UniRef50_D4MFL5: ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase domain
	UniRef50_S4ERW5: Putative bacteriocin-processing/bacteriocin ABC transporter, ATP-binding protein
	UniRef50_F0PCJ3: Antimicrobial peptide, streptococcin A-M57 family protein
	UniRef50_P36962: Bacteriocin lactococcin-G subunit beta
	UniRef50_D4MFL5: ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase domain
	UniRef50_F3R8B9: Type 2 lantibiotic biosynthesis protein LanM
Lactococcus lactis	UniRef50_G8P9Z6: Pleiotropic regulator of exopolysaccharide synthesis, competence and biofilm formation Ftr, XRE family
	UniRef50_Q9CHP4: Collagen adhesin
	UniRef50_Q07596: Nisin leader peptide-processing serine protease NisP
	UniRef50_Q9CJB7: Lactococcin A secretion protein LcnD-like
	UniRef50_D2BND8: Mucus-binding protein, LPXTG-anchored
	UniRef50_K7VR84: CHW repeat-/cell adhesion domain-containing protease and peptidase
	UniRef50_K7VVE3: Lactococcin A1
	UniRef50_P0A313: Bacteriocin lactococcin-A
	UniRef50_P0A3M8: Lactococcin-A immunity protein
	UniRef50_K7VRC5: Bacteriocin immunity protein A3
	UniRef50_P0A3M8: Lactococcin-A immunity protein
	UniRef50_P42708: Nisin immunity protein
	UniRef50_F8LI03: Sal9 lantibiotic transport ATP-binding protein
	UniRef50_P23648: Nisin-resistance protein
	UniRef50_P42708: Nisin immunity protein
	UniRef50_Q03202: Nisin biosynthesis protein NisC
Leuconostoc mesenteroides	UniRef50_B1MWF8: ABC-type metal ion transport system, periplasmic component/surface adhesin
	UniRef50_Q03VR9: ABC-type metal ion transport system, periplasmic component/surface adhesin
	UniRef50_Q03V11: Prebacteriocin
	UniRef50_C2KM58: Sodium bile acid symporter family protein (Fragment)
	UniRef50_C2KM58: Sodium bile acid symporter family protein (Fragment)
Serratia proteamaculans	UniRef50_V6A4U5: Fimbrial adhesin
	UniRef50_L0MJS0: Capsular exopolysaccharide biosynthesis protein
	UniRef50_P16316: Serralysin B
	UniRef50_Q11137: Serralysin

Table S3. Correlations between the relative abundances of microbial genera and the levels of volatile compounds

Genus	Compound(s)	R-value	Uncorrected p-value
Acetobacter	Acetic acid	0.76	<0.01
	2-methyl-1-butanol	0.65	0.01
	2,3-butanedione	0.67	<0.01
Kazachstania	Acetic acid	0.52	0.05
	2-methyl-1-butanol	0.53	0.04
	2,3-butanedione	0.85	<0.01
	2,3-pentanedione	0.68	<0.01
	2,3-hexanedione	0.72	<0.01
Lactobacillus	Carboxylic acids	0.6	0.02
	Esters	0.59	0.02
	3-methyl-1-butanol	0.58	0.02
Leuconostoc	2,3-butanedione	0.68	0.005
Saccharomyces	Carboxylic acids	0.71	<0.01
	Esters	0.78	<0.01

Table S4: Changes in the volatile profile of kefirs supplemented with (A) Lb. kefiranofaciens 484 NCFB 2797 and (B) L. mesenteroides DPC 7047.

	Compound	RT	CAS	LRI	Ref LRI	Non-spiked (%)	Spiked (%)	Difference (%)	Comment
Addition of *L. kefiranofaciens* to Fr1	3-Methylbutanol	6.992	123-51-3	728	**733**	0	0	0.00	NA
	2-Methylbutanol	7.058	137-32-6	731	**755**	2.65	2.39	-10.83	Addition of Lk resulted in a decrease in the % of 2-Methylbutanol
	Acetoin	7.188	513-86-0	737	**709**	64.25	56.90	-12.93	Addition of Lk resulted in a decrease in the % of Acetoin
	2,3-Butanediol	8.454	513-85-9	796	**802**	1.96	3.21	39.02	Addition of Lk resulted in an increase in the % of 2,3-Butanediol
	Acetic acid	5.938	64-19-7	667	**629**	10.28	12.50	17.79	Addition of Lk resulted in an increase in the % of Acetic acid
	2-Methylpropanoic acid	7.813	79-31-2	766	**774**	3.00	3.13	4.15	Addition of Lk resulted in an increase in the % of 2-Methylpropanoic acid
	2-Methyl-butanoic acid	10.25	116-53-0	863	**831**	1.09	1.17	6.32	Addition of Lk resulted in an increase in the % of 2-Methyl-butanoic acid
	Hexanoic acid	13.229	142-62-1	971	**983**	0.42	0.39	-6.74	Addition of Lk resulted in a slight decrease in the % of Hexanoic acid
	Octanoic acid	18.242	124-07-2	1156	**1160**	0.10	0.08	-17.83	Addition of Lk resulted in a slight decrease in the % of Octanoic acid
	Ethenyl acetate	4.683	108-05-4	557	**564**	0.78	1.91	59.15	Addition of Lk resulted in an increase in the % of Ethenyl acetate
	Ethyl acetate	4.929	141-78-6	587	**614**	0.00	0.12	100.00	Addition of Lk resulted in an increase in the % of Ethyl acetate
	Methyl-3-butyrate	10.071	503-74-2	856	**848**	6.45	8.82	26.83	Addition of Lk resulted in an increase in the % of Methyl-3-butyrate
	2-Methylbutyl acetate	10.492	624-41-9	872	**868**	0.41	0.47	11.44	Addition of Lk resulted in an increase in the % of 2-Methylbutyl acetate
	Acetone	3.975	67-64-1	470	**496**	2.83	3.29	13.87	Addition of Lk resulted in an increase in the % of Acetone
	2-Heptanone	10.875	110-43-0	886	**891**	0.08	0.22	65.86	Addition of Lk resulted in an increase in the % of Heptanone
Addition of *L. mesenteroides* to Idk	3-Methylbutanol	6.992	123-51-3	728	**733**	3.92	0.00	-100.00	Addition of Lm resulted in a decrease in the % of 3-Methylbutanol
	2-Methylbutanol	7.058	137-32-6	731	**755**	4.05	1.49	-63.22	Addition of Lm resulted in a decrease in the % of 2-Methylbutanol
	Acetoin	7.188	513-86-0	737	**709**	49.91	60.05	20.33	Addition of Lm resulted in an increase in the % of Acetoin
	2,3-Butanediol	8.454	513-85-9	796	**802**	2.34	2.68	14.91	Addition of Lm resulted in an increase in the % of 2,3-Butanediol
	Acetic acid	5.938	64-19-7	667	**629**	7.28	19.52	168.28	Addition of Lm resulted in an increase in the % of Acetic acid
	2-Methylpropanoic acid	7.813	79-31-2	766	**774**	1.38	3.01	117.76	Addition of Lm resulted in an increase in the % of 2-Methylpropanoic acid
	2-Methyl-butanoic acid	10.25	116-53-0	863	**831**	1.03	0.93	-9.97	Addition of Lk resulted in an increase in the % of 2-Methyl-butanoic acid
	Hexanoic acid	13.229	142-62-1	971	**983**	0.85	0.59	-31.50	Addition of Lm resulted in a decrease in the % of Hexanoic acid
	Octanoic acid	18.242	124-07-2	1156	**1160**	0.21	0.14	-32.72	Addition of Lm resulted in a decrease in the % of Octanoic acid
	Ethenyl acetate	4.683	108-05-4	557	**564**	1.08	0.70	-34.79	Addition of Lm resulted in a decrease in the % of Ethenyl acetate
	Ethyl acetate	4.929	141-78-6	587	**614**	0.00	0.06	0.00	Addition of Lk resulted in an increase in the % of Ethyl acetate
	Methyl-3-butyrate	10.071	503-74-2	856	**848**	8.01	7.21	-10.01	Addition of Lm resulted in a decrease in the % of Methyl-3-butyrate
	2-Methylbutyl acetate	10.492	624-41-9	872	**868**	0.23	0.46	99.70	Addition of Lk resulted in a decrease in the % of 2-Methylbutyl acetate
	Acetone	3.975	67-64-1	470	**496**	2.27	0.99	-56.43	Addition of Lm resulted in a decrease in the % of Acetone
	2-Heptanone	10.875	110-43-0	886	**891**	0.15	0.15	-1.52	Addition of Lm resulted in a decrease in the % of Heptanone

Table S5: Sensory terms for the ranking descriptive analysis of Kefir.

Attribute	Definition	Scale
Hedonic		
Appearance-Liking	The liking of appearance	0 = extremely dislike,10 = extremely like
Flavour-Liking	The liking of flavour	0 = extremely dislike,10 = extremely like
Aroma-Liking	The liking of aroma	0 = extremely dislike,10 = extremely like
Texture-Liking	The liking of texture	0 = extremely dislike,10 = extremely like
Overall acceptability	The acceptability of the product	0 = extremely unacceptable,10 = extremely acceptable
Intensity		
Kefir-like Flavour	Complex olfactory sensation due to fermentation of milk with kefir bacteria	0 = none, 10 = extreme
Medicinal Flavour	The flavours associated with Medicine	0 = none, 10 = extreme
Buttermilk Flavour	The flavours associated with Buttermilk	0 = none, 10 = extreme
Fruity/Estery flavour	The flavours associated with fatty acid ethyl esters	0 = none, 10 = extreme
Off-flavour	Off-flavour (Rancid)	0 = none, 10 = extreme
Bitter taste	Fundamental taste sensation of which caffeine or quinine in soda water is typical	0 = none, 10 = extreme
Sweet taste	Fundamental taste sensation of which sucrose is typical	0 = none, 10 = extreme
Sour	Fundamental taste sensation of which lactic acid is typical	0 = none, 10 = extreme
Prickling texture	A tingling feeling on the tongue similar to a carbonated mineral water	0 = none, 10 = extreme
Creamy texture	Velvet/soft feeling in the mouth (not fatty/oily)	0 = none, 10 = extreme
Mouth coating	Sensation of a thin film coating of the oral cavity	0 = none, 10 = extreme
Viscous texture	High resistance to flow in the mouth	0 = none, 10 = extreme
After taste	Lingering sour/milky taste associated with kefir	0 = none, 10 = extreme

Chapter 4

Omics-based insights into flavour development and microbial succession within surface-ripened cheese

Published in *mSystems*

(doi: https://doi.org/10.1128/mSystems.00211-17)

Authors: Aaron M. Walsh, Andrea S. Bertuzzi, J. J. Sheehan, Paul D. Cotter, Fiona Crispie, Paul L. H. McSweeney, Kieran N. Kilcawley, & Mary C. Rea. 2018.

Contributions:

- **Candidate** performed computational analysis
- **ASB** performed wet-lab work
- **JJS**, **PDC**, **FC**, **PLHM**, **KNK**, and **MCR** supervised the study

ABSTRACT

In this study, a young Cheddar curd was used to produce two types of surface-ripened cheese, using two commercial smear-culture mixes of yeasts and bacteria. Whole-metagenome shotgun sequencing was used to screen the microbial population within the smear-culture mixes, and on the cheese surface, comparing microorganisms both at species and strain level. The use of two smear mixes resulted in the development of distinct microbiota on the surface of the two test cheeses. In one case, most of the species inoculated on the cheese established themselves successfully on the surface during ripening; while in the other, some of species inoculated were not detected during ripening and the most dominant bacterial species, *Glutamicibacter arilaitensis,* was not a constituent of the culture mix. Generally, yeast species, such as *Debaryomyces hansenii* and *Geotrichum candidum,* were dominant during the first stage of ripening, but were overtaken by bacterial species, such as *Brevibacterium linens* and *G. arilaitensis,* in the later stages. Using correlation analysis, it was possible to associate individual microorganisms with volatile compounds detected by GC-MS in the cheese surface. Specifically, *D. hansenii* correlated with the production of alcohols and carboxylic acids, *G. arilaitensis* with alcohols, carboxylic acids and ketones and *B. linens* and *G. candidum* with sulphur compounds. In addition, metagenomic sequencing was used to analyse the metabolic potential of the microbial population on the surface of the test cheeses, revealing a high relative abundance of metagenomic clusters associated with the modification of colour, variation of pH and flavour development.

INTRODUCTION

Recent studies, utilising metagenomics alongside metabolomics, have begun to address the role of the microbiota in the biochemical dynamics of fermentation processes (1-4). It is clear that in fermented foods, the metabolic interactions which regulate the composition of the microbial population influence the taste, shelf life and safety of the subsequent product (5). The ability to manipulate fermented food microbiota represents an important avenue for the food industry to develop new food products with precise characteristics.

Surface-ripened cheese, such as Münster, Tilsit, Livarot, Limburger or Comté, is characterised by the growth of a heterogeneous microbiota on the cheese surface, with the consequent development of a strong flavour. The flavour and the appearance of these types of cheese are related to the metabolic activities of bacteria and yeasts, which comprise the smear consortium. Generally, the cheese is brined or surface-salted, which also influences the growth of surface microbiota. In some traditional procedures, young cheese is smeared by transferring the smear from older cheese to younger curd ("old-young" technique) (6, 7). However, today, commercial mixtures of smear bacteria and yeasts are more commonly used to produce a more standardised product.

Metagenomic sequencing has proven to be a valid method for investigating the microbial population on the exterior of the surface-ripened cheese (3, 8-10). In studies of complex microbial communities in fermented foods, such as kefir, the information gained through whole-metagenome shotgun sequencing allows variations of the microbial populations, and also the metabolic pathways involved in the fermentation processes, to be monitored (1).

The aim of the current study was to investigate, at both the species and strain levels, the succession of the microbial populations present on the rind of a surface-ripened cheese, produced with young Cheddar cheese curd as a base, using two different commercial smear-culture mixes. Studies were performed over the course of 30 days of ripening to correlate volatile analysis with data generated through whole-metagenome shotgun sequencing in order to understand how microbial composition related to flavour development. Moreover, metagenomic analysis allowed for the screening of metagenomic clusters during cheese ripening, showing the involvement of the surface microbiota in a variety of biochemical processes.

MATERIAL AND METHODS

Smearing of cheese blocks

A block of commercial Cheddar cheese, < 24 hours after manufacture, was aseptically cut into smaller blocks (~8 × 6.5 × 30 cm) and washed with smearing solutions, as described in our previous study (11). Two commercial smear-culture mixes comprising of *Geotrichum candidum*, *Debaryomyces hansenii*, *Brevibacterium linens*, *Glutamicibacter arilaitensis* and *Staphylococcus xylosus* (S5 mix) (Sacco, Cadorago, Italy) and *D. hansenii*, *Cyberlindnera jadinii*, *Brevibacterium casei* and *B. linens* (D4 mix) (DuPont™ Danisco®, Beaminster, Dorset, UK) were used to inoculate the surface of the cheese curd. The blocks of cheese were washed with the smearing solutions and placed in sterile racks inside a sterile plastic bags (Südpack Verpackungen, Ochsenhausen, Germany), as previously described (11). The cheese was ripened for 30 days at 15°C, with a relative humidity of ~97%. At days 7, 10 and 15 of ripening, the cheese blocks were brushed with a sterile sponge, soaked in a sterile brine solution (5% NaCl), to uniformly spread the smear microbiota on the cheese

surface. As a control, un-smeared cheese blocks were vacuum-packed in sterile bags and incubated at 15°C, similarly to the test cheeses.

Sampling cheese

Three replicate cheese trials were performed at different times during Cheddar cheese making season. All data presented are the results of the analysis performed on samples taken from the cheese surface (at a depth of ~ 0.5cm). All analyses were performed in triplicate.

Compositional analysis and pH

pH level was measured on day 0, 18, 24 and 30 using a standard pH meter (Mettler-Toledo MP220, Schwerzenbach, Switzerland). The data were analysed by one-way analysis of variance (ANOVA) using SAS 9.3 (12).

Determination of colour

At day 0, 18, 24 and 30 of ripening, the colour was measured on the cheese surface at room temperature, using a Minolta Colorimeter CR-300 (Minolta Camera, Osaka, Japan). The instrument was calibrated on white tile, and the colour of the cheese surface was measured using L^*, a^* and b^*-values. L^* value measures the visual lightness (as values increase from 0 to 100), a^* value measures from the redness to greenness (positive to negative values, respectively) and b^* value from the yellowness to blueness (positive to negative values, respectively).

Total DNA extraction from cheese surface

The total DNA was extracted from the smear culture mixes and the cheese samples using the PowerSoil DNA Isolation kit, as described in the manufacturer's protocol (Cambio, Cambridge, United Kingdom). For the DNA extraction from the cheese surface, at day 0, 18, 24 and 30, a pre-treatment step was included as follows. Samples were removed from different parts of the cheese block and pooled to give a representative sample of 5 g. The cheese was placed in a stomacher bag with 50 ml of 2% trisodium citrate and homogenised using a masticator mixer (IUL S.A., Barcellona, Spain) for 5 min.

15 ml of the smear-culture mix, or the cheese solution, were placed into sterile falcon tubes and centrifuged for 30 min at 4,500×g. After centrifugation, the supernatant was discarded and the pellet was placed in a 2 ml Eppendorf tube. The pellet was washed several times with sterile phosphate buffered saline (PBS) by centrifuging at 14,500×g for 1 min, until the supernatant was completely clear. The pellet was then added to PowerBead tubes (Cambio, Cambridge, United Kingdom) provided with the kit as described in the protocol and homogenised by shaking on the TissueLyser II (Qiagen, West Sussex, United Kingdom) at 20 Hz for 10 min. The DNA was then purified according to the protocol of the standard PowerSoil DNA Isolation kit (Cambio, Cambridge, United Kingdom).

Total DNA was initially qualified and quantified by gel electrophoresis and the NanoDrop 1000 (BioSciences, Dublin, Ireland) before more accurate quantification with Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland).

Whole-metagenome shotgun sequencing

Whole-metagenome shotgun libraries were prepared in accordance with the Nextera XT DNA Library Preparation Guide from Illumina (13). Libraries for the starter mixture samples were sequenced on the Illumina MiSeq, with a 2 x 300 cycle v3 kit. Libraries for the cheese samples were sequenced on the Illumina NextSeq 500, with a NextSeq 500/550 High Output Reagent Kit v2 (300 cycles). All sequencing was done in the Teagasc sequencing facility, in accordance with standard Illumina sequencing protocols.

Bioinformatic analysis

Raw whole-metagenome shotgun sequencing reads were processed, on the basis of quality and quantity, using a combination of Picard Tools (https://github.com/broadinstitute/picard) and SAMtools (14). Processing of raw sequence data produced a total of 3,214,480 ± 841,719 filtered reads for samples sequenced on the MiSeq, and 19,210,475 ± 12,478,696 filtered reads for samples sequenced on the NextSeq. The metagenomic binning tool Kaiju (15) was used to determine the species-level microbial compositional of samples. The NCBI non-redundant protein database (16) was used with Kaiju. PanPhlAn (17) was used for strain-level analysis of species of interest. PanPhlAn works by aligning sequencing reads against a species pangenome database, built from reference genomes, to identify the gene families present in strains from metagenomic samples. The reference genomes included for each pangenome database are outlined in Table S1. SUPER-FOCUS (18) was used to characterise the microbial metabolic potential of samples. SUPER-FOCUS measures the abundances of subsystems, or groups of proteins with shared functionality, by aligning sequencing reads against a reduced SEED (19)

database. Sequencing reads have been deposited in the European Nucleotide Archive under the project accession number PRJEB15423.

Free amino acid analysis

FAA analysis was performed at the end of the ripening (day 30) on the soluble N extracts using a Jeol JLC-500V AA analyser fitted with a Jeol Na+ high performance cation exchange column (Jeol Ltd., Garden city, Herts, UK) (20). The chromatographic analyses were conducted at pH 2.2. Results were expressed as µg mg-1 of cheese.

Free fatty acid analysis

FFA extracts were performed at the end of the ripening (day 30) according to the method outlined by De Jong and Badings (21). The FFA extracts were derivitised as methyl esters as described by Mannion et al. (22). Fatty acid methyl esters extracts were analysed using Varian CP3800 gas chromatograph (Aquilant, Dublin 22, Ireland) with a CP84000 auto-sampler and flame ionisation detector and a Varian 1079 injector (Aquilant, Dublin 22, Ireland). Results were expressed as µg mg-1 of cheese.

Volatile analysis

The volatile compounds were analysed at days 0, 18, 24 and 30. The surface of the cheese was removed, wrapped in foil and stored vacuum-packed at -20°C until analysis. Before analysis the samples were defrosted, grated and 4 g of cheese surface were used. Analysis was carried out as outlined by Bertuzzi *et al.* (11).

Statistical analysis

Statistical analysis was done with SAS 9.3 (12) and R-3.2.2 (23). The R packages ggplot2 and pheatmap were used for data visualization. The vegan package was used to calculate the Bray-Curtis dissimilarity between samples, while the Hmisc package was used for correlation analysis.

RESULTS

Microbial composition of the smear-culture mixes

Two smear-culture mixes D4 and S5 were used for the cheese trials, and contained, as outlined in the supplier specification sheet, *Brevibacterium linens*, *Debaryomyces hansenii*, *Cyberlindnera jadinii* and *Brevibacterium casei*, or *Staphylococcus xylosus*, *B. linens*, *D. hansenii*, *Geotrichum candidum* and *Glutamicibacter arilaitensis* (previously *Arthrobacter arilaitensis*), respectively. Using metagenomic analysis, performed with Kaiju, the relative abundances of the individual species within the mixes were determined (Figure 1). Overall, Kaiju was able to assign 81.7 ± 1.5% of reads from the starter mix samples at the species-level. The proportion of assigned reads for each starter mixture sample is presented in Figure S1. *B. casei* (60.83%) and *C. jadinii* (15%) were the most abundant bacterial and yeasts species in D4, while *B. linens* and *D. hansenii* were minor components in the smear-culture mix with relative abundances of 5.25% and 1.92%, respectively (Figure 1; Table S1). In the S5 mix, *G. arilaitensis* (64.25%) together with *D. hansenii* (14.56%) and *G. candidum* (11.83%) were the most abundant bacterial and yeasts; *S. xylosus* (0.59%) and *B. linens* (3.52%) were present at lower relative abundances. Other species, not specified by the suppliers, were identified at low relative abundance in the smear-culture mixes D4 and S5, and are reported in Table S1.

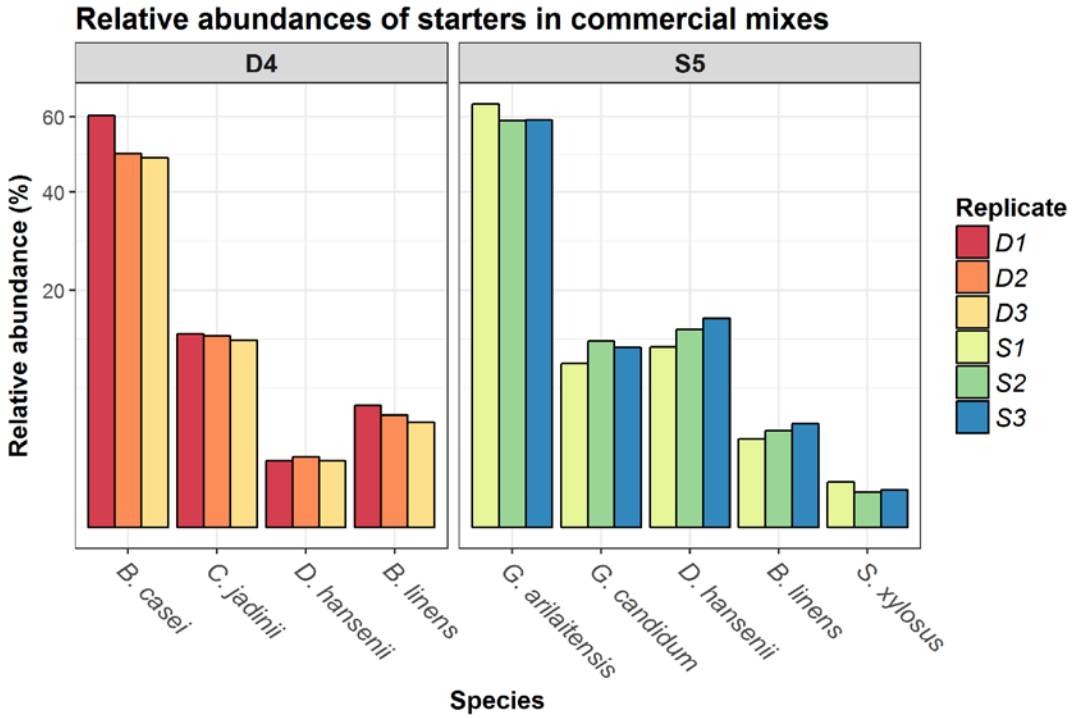


Figure 1: Relative abundances of the species (%), which were indicated as being present by the supplier, within the smear-culture mixes D4 and S5 (replicates of three analyses DA, DB, DC, and SA, SB, SC).

Species-level composition of the cheese surface

Two test cheeses, D4 and S5, were prepared by smearing young Cheddar cheese curd with the two aforementioned commercial smear-culture mixes and ripened for 30 days at 15°C. Kaiju was used to determine the bacterial and yeast composition of the cheese surface at day 0, 18, 24 and 30, for both the control cheese (un-smeared and ripened under vacuum) and the two test cheeses. Overall, Kaiju was able to assign 57.5 ± 8.3% of reads from the cheese samples at the species-level. The proportions of assigned reads for each sample are presented in Figure S2. Compositional data of the cheese surface were analysed by a one-way analysis of variance (ANOVA), designed with SAS 9.3 to determine the significant differences in the proportions of the individual species present over time. As expected, lactic acid bacteria dominated the surface of all samples at day 0, and their relative abundance on the surface of the control did not significantly change throughout the 30 days of ripening (Figure 2). *L. lactis* and *S. thermophilus* were identified in all samples analysed (D4, S5 and control) (Figure 2). *L. lactis* was the dominant species in the control, constituting 75.85% of the initial population at day 0, decreasing to 65.99% at day 30. *S. thermophilus* increased from 19.65% at day 0 to 28.21% at day 30, while the relative abundance of *Lb. helveticus* was low throughout the ripening period (2.12% at day 0, and 2.72% at day 30) (Table S2). However, over the course of 30 days of ripening, the smearing processes clearly influenced the microbial population of the cheese surface of both test cheeses, D4 and S5, causing a significant reduction in the relative abundance of *Lb. helveticus* ($P<0.03$) and *L. lactis* ($P<0.0001$). From day 0 to day 18, the population on the surface of D4 changed from predominately LAB to *Debaryomyces hansenii* and *Glutamicibacter arilaitensis* (Figure 2). Subsequently, over the course of ripening, the relative

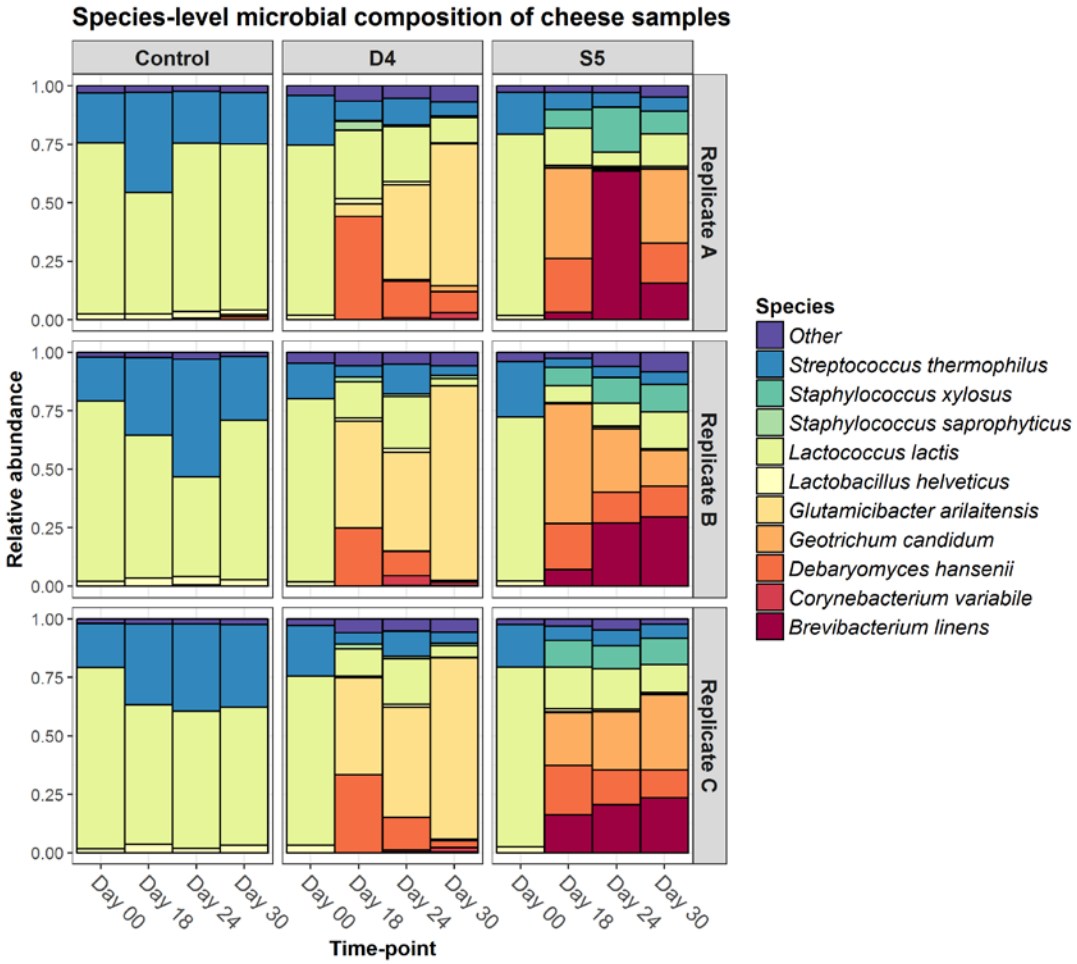


Figure 2: Relative abundance at the species-level of the microbiota on the cheese surface of control, D4 and S5 at day 0, 18, 24 and 30. Data shown for the three replicate trials.

abundance of *D. hansenii*, significantly decreased (*P<0.0001*) from 34.12% at day 18 to 4.14% at day 30 (Table S2). In parallel, the relative abundance of *G. arilaitensis* significantly increased (*P<0.0001*) from 30.9% at day 18, to become the dominant population on the cheese surface (73.75%) at day 30 (Table S2). The secondary microbial population (individually between 1% and 3% of the population) of the D4 surface was composed of species not included in the initial smear-culture mix, and included *Arthrobacter sp., Corynebacterium variabile, Debaryomyces fabryi, G. candidum, Staphylococcus equorum* and *Staphylococcus saprophyticus* (Table S2). In addition, some species present in the initial smear-culture mix (*C. jadinii* and *B. casei*) were not detected during ripening, while *B. linens* was detected at only at a very low relative abundance on the cheese surface of D4 throughout ripening (Table S2). By comparison, the microbiota was more diverse in cheese S5 (Figure2; Table S2). On the cheese surface of S5, the relative abundance of the LAB decreased, while *B. linens* increased significantly (*P<0.004*) from day 18 to day 24, reaching 37.05 %, before decreasing, but not significantly, to 22.84% at day 30 (Table S2). The yeasts *D. hansenii* and *G. candidum* (components of the S5 mix) were the most abundant population on the cheese surface at day 18, comprising 21.2% and 37.54% of the microbiota, respectively, but their relative abundance significantly decreased (*P<0.04*) by day 24 to 9.57% and 17.6%, respectively, without showing further significant reductions at day 30 (Table S2). *S. xylosus* was detected at 9.08% at day 18, and did not change significantly throughout the ripening period (Table S2). In addition, a secondary microbial population, comprising of *D. fabryi* (detected in the S5 mix; Table S1) and *Psychrobacter sp* (not detected in the S5 mix; Table S1), developed at low relative abundance (1-2%) on the surface of the cheese S5 (Table S2) over the course of the ripening period. However, some

inoculated species were either not detected (*S. equorum*) at any stage throughout ripening, or detected at very low relative abundance (*G. arilaitensis* ~0.44%) on the cheese surface during ripening (Table S2).

Strain-level analysis of bacterial starter and smearing cultures

The metagenomic sequences of the bacteria used as starter cultures in the Cheddar cheese curd (*L. lactis* and *S. thermophilus*) and as smearing cultures (*B. linens*, *S. xylosus*, and *G. arilaitensis*) were compared at the strain-level, using PanPhlAn, to determine the presence/absence of the inoculated bacterial strains on the cheese throughout ripening and to investigate possible cross-contamination between the samples. PanPhlAn indicated that the same *L. lactis* and *S. thermophilus* strains were present in each cheese throughout ripening (Figure 3). In contrast, the *B. linens* strains detected in all D4 samples appeared to be distinct from those in all S5 samples (Figure 3). Additionally, the *B. linens* strains detected on both cheeses clustered with those present in their respective starter cultures (Figure 3), which suggests that the inoculated *B. linens* strains colonised the cheese surfaces. As mentioned, although *G. arilaitensis* was present in the D4 mix, but not the S5 mix, this species was only detected on S5 cheeses, which suggested possible cross-contamination between the samples. However, PanPhlAn indicated that these *G. arilaitensis* strains were distinct (Figure 3). Interestingly, though, the *G. arilaitensis* strain from the D4 mix did appear to cluster more closely to that detected on the control cheeses (Figure 3). Finally, the *S. xylosus* strain from the S5 mix was distinct from that present on the S5 cheeses, which suggests that the inoculated strain may not have colonised the cheese surface.

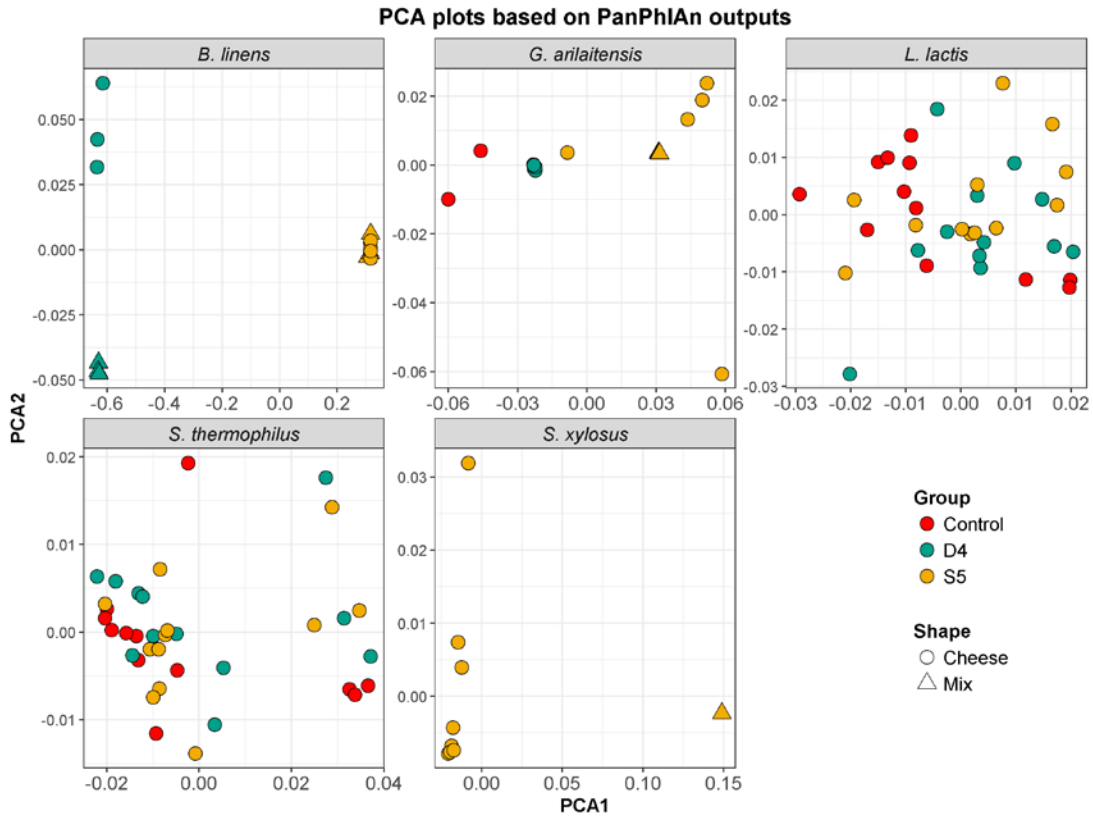


Figure 3: Principal-component analysis (PCA) plot of the profiles of the strains determined by PanPhlAn.

Volatile compounds present on the cheese surface

Headspace solid phase micro-extraction (HS-SPME) gas chromatography-mass spectrometry (GC-MS) was used to analyse the development of volatile compounds at day 0, 18, 24 and 30 of ripening, for both control and test cheeses. In total, 53 volatile compounds that could potentially contribute to the flavour development were detected on the cheese surface. These compounds are predicted to arise from a variety of substrates, and consisted of 8 alcohols, 6 aldehydes, 10 carboxylic acids, 10 esters, 13 ketones, 2 *S*-thioesters and 4 sulphur compounds (i.e. a total of 53 compounds). As expected, given the microbial diversity on the surface there was a greater variety and intensity of volatile compounds detected compared to the control cheese, in which only 23 of the aforementioned 53 compounds were detected. In all cheeses, all volatiles were detected increased throughout the ripening period, apart from 2,3-butanediol, hexanal, heptanal, octanal, nonanal, 2,3-butanedione and dimethylsulfone.

Correlations between microbial taxa and volatile compounds

Correlation analysis on the relative abundance of microbial species and the abundance of volatile compounds detected on the cheese surface was performed using the Spearman correlation test, as described previously by Walsh *et al.* (1). From the results of the metagenomic analysis (performed with Kaiju) and the volatile analysis, it was possible to associate both yeasts and bacteria, at species-level, with specific volatile compounds. Figure 4 demonstrates the degree of correlation between the volatile compounds and the organisms detected.

There was a strong correlation between *B. linens* and *G. candidum* with sulphur compounds and 2-methyl-1-butanol; *S. xylosus* was correlated with sulphur

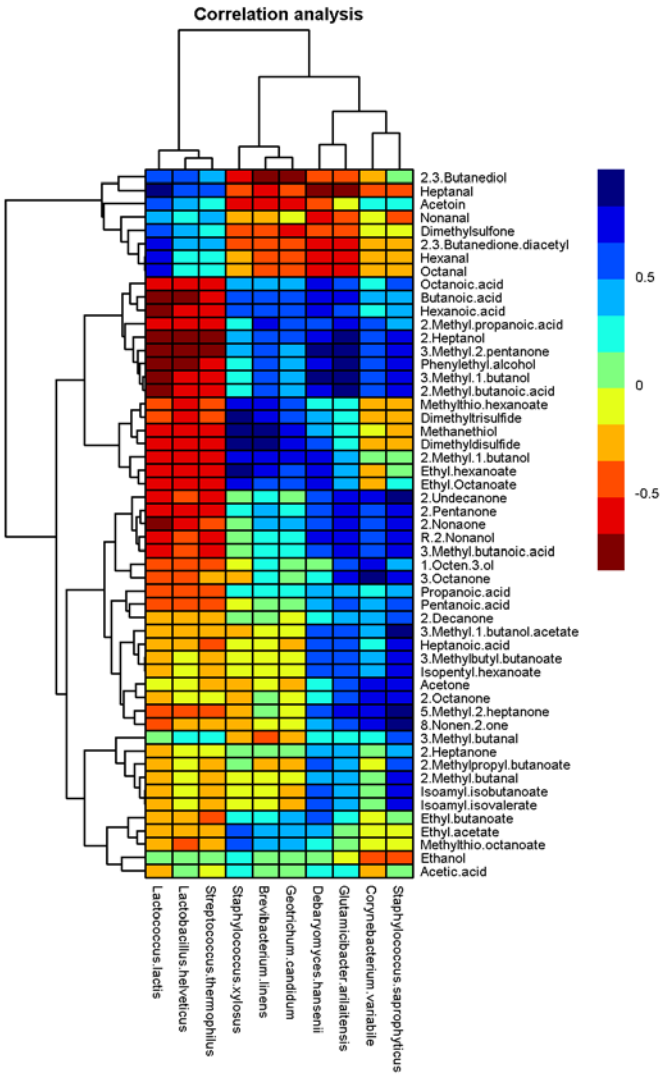


Figure 4: Hierarchically clustered map showing the correlation between the relative abundance of the microbial species and the levels of volatile compounds detected on the cheese surface. Clustering was performed by using the hclust function in R. The colour of each tile of the heat map indicates the level of correlation for a given species-compound combination, as indicated by the colour key.

compounds, 2-methyl-1-butanol and some ethyl esters; *C. variablile* was correlated with ketones; *D. hansenii* was correlated with acids and alcohols; *G. arilaitensis* was correlated with ketones and alcohols and acids, and *S. saprophyticus* with ketones, esters, acids and alcohols (Figure 4; Table 1).

Gene content of cheese surface microbiota

Using SUPER-FOCUS, whole-metagenome shotgun sequencing was used to characterise the functional potential of the whole microbial community on the cheese surface at different stages of ripening. Overall, SUPER-FOCUS was able to assign 62.5 ± 10.9% of reads from the cheese samples to a function. The proportions of assigned reads for each cheese sample are presented in Figure S2. The functional clusters analysed were initially organised into three different levels, in relation to the specificity of the metabolic pathways. Pathway data was analysed to determine the significant differences of the individual metabolic clusters by one-way analysis of variance (ANOVA), using SAS 9.3, with the selection of sixteen specific functional clusters with relative abundance significantly higher ($P<0.05$) on the cheese surface of S5 and D4, compared to the control (Figure 5).

Colour and pH variation

pH and colour analysis was performed on the three cheese types and the resultant data was examined using a split-plot test, designed with SAS 9.3. A significant interactive effect ($P<0.0001$) between smear treatments and ripening time was observed for pH. At days 18, 24 and 30, the pH was significantly higher ($P<0.0001$) on the surface of S5 and D4, compared to the control. In addition, the pH was

Table 1: List of strong positive correlations (R>+0.5) between the levels of volatile compounds and the relative abundance of species on the cheese surface.

Correlation species and compound	Potential precursor	R value
Debaryomyces hansenii		
2-Methyl butanoic acid	Isoleucine	0.81
3-Methyl-1-butanol	Leucine	0.85
Octanoic acid	Lipolysis	0.76
Hexanoic acid	Lipolysis	0.81
2-Heptanol	2-Heptanone (fatty acid oxidation)	0.8
Glutamicibacter arilaitensis		
2-Methyl butanoic acid	Isoleucine	0.9
3-Methyl-1-butanol	Leucine	0.86
3-Methyl butanoic acid	Leucine	0.77
Phenylethyl alcohol	Phenylalanine	0.83
3-Methyl-2-pentanone	Fatty acid oxidation	0.89
2-Undecanone	Fatty acid oxidation	0.82
5-Methyl-2-heptanone	Fatty acid oxidation	0.78
2-Pentanone	Fatty acid oxidation	0.77
2-Nonaone	Fatty acid oxidation	0.76
2-Heptanol	2-Heptanone (fatty acid oxidation)	0.86
Geotrichum candidum		
2-Methyl-1-butanol	Isoleucine	0.76
Methanethiol	Methionine	0.76
Dimethyldisulfide	Methanethiol	0.79
Brevibacterium linens		
2-Methyl-1-butanol	Isoleucine	0.81
Methanethiol	Methionine	0.82
Dimethyldisulfide	Methanethiol	0.85
Dimethyltrisulfide	Methanethiol	0.77
Staphylococcus xylosus		
2-Methyl-1-butanol	Isoleucine	0.77
Methanethiol	Methionine	0.84
Dimethyldisulfide	Methanethiol	0.95
Dimethyltrisulfide	Methanethiol	0.86
Methylthio hexanoate	Methanethiol hexanoic acid	0.78
Ethyl hexanoate	Ethanol hexanoic acid	0.85
Ethyl octanoate	Ethanol octanoic acid	0.77
Staphylococcus saprophyticus		
2-Methyl-butanoic acid	Isoleucine	0.76
3-Methyl-1-butanol	Leucine	0.77
Heptanoic acid	Lipolysis	0.76
5-Methyl-2-heptanone	Fatty acid oxidation	0.98
2-Undecanone	Fatty acid oxidation	0.88
8-Nonen-2-one	Fatty acid oxidation	0.87
3-Methyl-2-pentanone	Fatty acid oxidation	0.77
2-Nonanol	2-Nonaone (fatty acid oxidation)	0.78
Isopentyl acetate	3-Methyl-1-butanol acetic acid	0.87
Isopentyl butanoate	3-Methyl-1-butanol butanoic acid	0.8
Isopentyl hexanoate	3-Methyl-1-butanol hexanoic acid	0.8
Corynebacterium variabile		
3-Octanone	Fatty acid oxidation	0.99
2-Octanone	Fatty acid oxidation	0.78
5-Methyl-2-heptanone	Fatty acid oxidation	0.77

[a] Correlations for which the *P* value was 0.001 (corrected for multiple comparisons using the Bonferroni method) and the *R* value was 0.75.

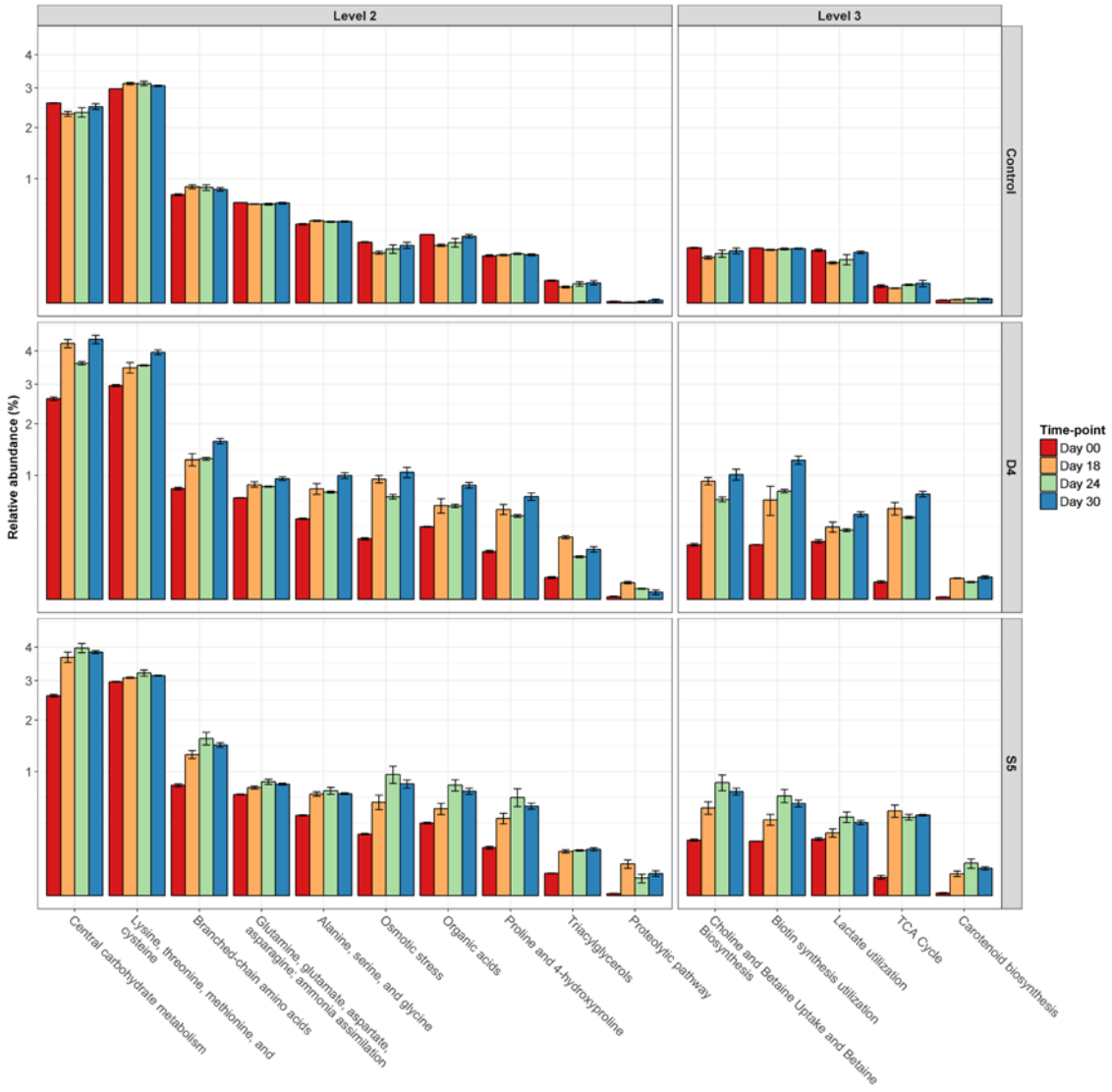


Figure 5: Average and standard error (SE) between the three replicate trials of the relative abundance of significantly different (P<0.05) metagenomic clusters detected with SUPER-FOCUS at day 0 (red), 18 (orange), day 24 (green) and 30 (blue), for the cheese surface of control, D4 and S5.

significantly higher (*P<0.0001*) on the surface of S5, compared to D4, from day 18 onwards (Figure S3).

A significant interactive effect (*P<0.0001*), between time and smear treatments, was observed on L^*, a^* and b^* values. At days 18, 24 and 30, the a^* value was significantly higher (*P<0.0001*) for the surface of S5 and D4, compared to the control. At day 30, the a^* value was also significantly higher (*P<0.02*) on the surface of D4 compared to S5 (Figure S4).

Free amino acids and fatty acids

Free amino acid (FAA) and free fatty acid (FFA) analysis was performed on the three cheese types and the experimental results were examined by one-way analysis of variance (ANOVA), using SAS 9.3. The concentrations of total FAAs on the surface of S5 (15158±1683 μg mg^{-1}) and D4 (11914±1769 μg mg^{-1}) were significantly higher (*P<0.05*) than those on the control surface (6605±819 μg mg^{-1}). In addition, some individual FAAs, such as tyrosine, proline and histidine, were significantly higher (*P<0.05*) on the surface of S5, compared to the surface of D4 and the control (Figure S5).

The concentrations of total FFAs on the surface of S5 (22069±3875 μg mg^{-1}) and D4 (26562±2606 μg mg^{-1}) were significantly higher (*P<0.05*) compared to the control (1336±70 μg mg^{-1}). Some individual FFAs, such as C4:0, C8:0, C10:0, C12:0, C14:0 and C18:0, were significantly higher (*P<0.05*) on the surface of D4, compared to S5 or the control (Figure S5).

DISCUSSION

In this study, the use of whole-metagenome shotgun sequencing facilitated the characterisation, at species and strain levels, of microbial succession among smear microorganisms (both bacteria and yeasts) on the cheese surface, and the analysis of the metabolic potential of the whole microbial community at different stages of ripening. Volatile flavour compounds were analysed over time, using HS-SPME GC-MS, and correlated with the microbial species that developed during ripening.

Cheddar cheese curd, < 24h post manufacture, was inoculated with two different smear-culture mixes and incubated at 15°C, for 30 days. Un-smeared Cheddar cheese curd, vacuum packed to prevent the growth of spoilage moulds on the cheese surface, was used as a control. This model was chosen to investigate the microbial succession and flavour development as it had been shown in a previous study that yeasts and bacteria establish themselves satisfactorily on the surface of young Cheddar cheese curd, producing cheese with modified flavour and appearance (11).

On the cheese surface of S5 and D4, a very heterogeneous microbial consortium developed during ripening, triggering an array of biochemical processes. Yeasts are considered the responsible of the deacidification of the cheese surface (observed on S5 and D4; Figure S3) by the degradation of lactate (to CO_2 and H_2O) (24, 25) together with the formation of alkaline metabolites (from metabolism of FAAs) (26), and the secretion of growth factors (vitamins and amino acids) which support the growth of bacteria (25, 27). As expected, in parallel with the growth of the yeasts, the relative abundance of the metagenomic clusters related to lactate- utilization, and the biosynthesis and uptake of biotin, was higher for the cheese surface of D4 and S5, compared to the control (Figure 5). During ripening, the surfaces of D4 and S5 were washed with a 5% salt solution, causing a hyperosmotic stress on the microbial population of the cheese surface (28). This correlated with a higher relative abundance

for the metagenomic clusters related to osmotic stress resistance and metabolism of choline and betaine (osmoprotectants) (29), for the washed cheeses compared to the unwashed control (Figure 5).

The development of a red/orange colour on the surface is an important characteristic of many smear ripened cheeses. This colour development is usually derived through the metabolism of carotenoids (30, 31), and correspondingly higher relative abundance of metagenomic clusters, involved in the metabolism of the carotenoids (carotenoids and carotenoid biosynthesis), was observed on the surface of the cheese S5 and D4, compared to the control (Figure 5).

Surface-ripened cheeses are also characterised by a strong flavour which is driven by the biochemical metabolism of the microbial consortium which develops on the cheese surface over time. These are associated with proteolytic and lipolytic pathways, driving the increase in the levels of FAAs and FFAs. These pathways, together with lactose and citrate metabolism, are considered to be responsible for the main precursors of flavour compounds in cheese. In the current study, the relative abundance of the metagenomic clusters associated with the proteolytic pathway and the metabolism of triacylglycerols was higher for D4 and S5, compared to the control, which was consistent with FAA- and FFA-related data (Figure S5). During ripening, the relative abundance of metagenomic clusters directly related to the formation of volatile compounds, such as carbohydrate metabolism, organic acids (including FFAs) and FAAs (except aromatic amino acids), and indirectly related, such as TCA cycle (important for the α-ketoglutarate production), was significantly higher ($P<0.05$) for the cheese surface of both D4 and S5, compared to the control (Figure 5). Correspondingly, numerous volatile compounds (alcohols, aldehydes, carboxylic acids, ketones, sulphur compounds, esters and *S*-thioesters) (Figure 3) were produced

on the cheese surface of S5 and/or D4, conferring an intense flavour to the cheese surface of D4 and S5.

During ripening, on the cheese surface of S5 and D4, a microbial succession, involving various inoculated, and indeed some non-inoculated, microorganisms, was apparent. Consistent with other studies, specific smear strains, added as adjunct cultures to the milk, or on the exterior of surface-ripened cheese during manufacture, have not been detected at the end of ripening (32-36). In this study, the species detected on the cheese surface by metagenomic analysis did not fully correspond with the components of the smear-culture mixes. Different contaminant populations developed on the cheese surface of both test cheeses, especially on D4, probably due to the different interactions and competition with the cultures of the two mixes (Figure 2; Table S2).

D. hansenii was part of the inoculum used for both S5, and D4 surface. *D. hansenii* is a component of the surface microbiota of many surface-ripened cheeses, and is very tolerant to high salt concentrations and low pH conditions (24, 37). Presumably due to these characteristics, *D. hansenii* was present at high relative abundance in both test cheeses, mainly in the early stage of ripening (at day 18), and then decreased gradually in the later stages (day 24 and 30) (Table S2). Volatile compounds significantly associated ($P<0.001$) with *D. hansenii* were mainly alcohols and carboxylic acids (Figure 4; Table 1). The biosynthesis of branched chain alcohols and carboxylic acids, from FAA metabolism, and the biosynthesis of medium-long carboxylic acids, from FFA metabolism, are processes mainly attributed to yeast and mould metabolism, including *D. hansenii* (38-43).

On the cheese D4, the relative reduction of *D. hansenii* with time, corresponded to an increase of Gram-positive bacteria. *G. arilaitensis,* a component of S5 mix, did not grow on the cheese surface of S5, and, though not inoculated as part of the culture

mix, was the dominant bacteria on the surface of D4 (Figure 2; Table S2). Through the use of PanPhlAn, which uses metagenomic data to achieve strain-level microbial profiling resolution, we have demonstrated that the *G. arilaitensis* strain, present on D4, was not the same strain as inoculated onto S5 (Figure 3). The inability of the inoculated *G. arilaitensis* strain to grow on the S5 cheese is most likely due to the different interactions within the microbiota on the cheese surface. Other studies on the microbial composition of the surface of Limburger cheese observed that *G. arilaitensis* behaved in a similar manner, showing high relative abundance when it was co-inoculated only with *D. hansenii*, while showing low relative abundance when combined with both *D. hansenii* and *G. candidum* (25). That *G. arilaitensis* contributes to cheese flavour has been shown previously in model cheese media (44) (producing alcohols, and especially ketones), and in the current study, where it was significantly (*P<0.001*) associated with 3-methyl-1-butanol and phenylethyl alcohol, branched carboxylic acids (from FAAs metabolism), 2-heptanol and ketones (from FFAs metabolism) (Figure 4; Table 1). In addition a genomic study showed numerous genes encoding for protein degradation and fatty acid oxidation in *G. arilaitensis* (45).

On the cheese surface of S5, *G. candidum* was co-inoculated with *D. hansenii* and established itself to become the most abundant yeast population by day 18. The successful cohabitation of *G. candidum* and *D. hansenii* may be explained by the fact that they do not compete for energy sources in the same way in cheese. *D. hansenii* uses lactate, or the limited amount of lactose present in the cheese post manufacture (0.8-1%), while *G. candidum* preferentially uses only lactate (29, 46). During ripening, sulphur compounds were significantly associated (*P<0.001*) with *G. candidum* (Figure 4; Table 1), which is in agreement with other studies which have shown that

G. candidum is able to catabolise methionine in one-step degradation, with the biosynthesis of sulphur compounds (42, 47, 48).

The production of sulphur compounds is an important characteristic of many surface ripened cheese and *B. linens* is considered one of the main species responsible for the development of the strong flavour of many surface-ripened cheese through the biosynthesis of sulphur compounds derived from methanethiol. In this study *B. linens* was present at relatively low abundance in the original culture mixes (5.26% and 3.53% for D4 and S5, respectively; Table S1). However, while detected at very low relative abundance on the cheese surface of D4, was one the most dominant bacteria detected on S5 (37.05% at day 24; Table S2). While this may be due to inter-strain differences, it is most likely due to the different interactions within the microbiota of S5 and D4. Studies have shown that *B. linens* does not always establish itself on the cheese surface during ripening, even if it is present in the initial culture mix (33-35, 49, 50). However, in previous studies *G. candidum* has been shown to stimulate the growth of *B. linens* in co-culture (51), suggesting the hypothesis that in S5, *G. candidum*, present at high relative abundance, might have likely produced growth factors that supported the growth of *B. linens*; while in D4, it was out-competed by *G. arilaitensis*, which established itself very quickly on the surface of S5 and made up 75% of the microbiota at the end of ripening. *B. linens* was significantly associated ($P<0.001$) with methanethiol and its derivatives (dimethyldisulfide and dimethyltrisulfide) (Figure 4; Table 1), which likely originated from the one-step degradation of methionine (38, 44, 52, 53).

Other species, while present at lower relative abundance on the cheese surface of S5 and D4, were also responsible for the biosynthesis of some volatile compounds. *S. xylosus*, present in the S5 mix, was not as successful as *B. linens* at establishing itself

on the cheese surface, and was present at only at 10.83-13.36% of relative abundance, during ripening (Table S2). This is most likely due to competition for nutrients within the microbiota, as suggested by Mounier et al. (46). Members of the genus *Staphylococcus* can establish themselves on surface ripened cheese in the early stages of ripening but are regularly overtaken by other bacteria at the later stages (34, 54, 55).

In this study, specific species detected in low relative abundances in S5, such as *S. xylosus* (9.08-13.36%), and in D4, such as *S. saprophyticus* (1.06-2.69%), and *C. variable* (2.04-2.08%) (Table S2), were significantly associated ($P<0.001$) with a range of flavour compounds important in surface-ripened cheese (Figure 4; Table 1), and interestingly, while *S. xylosus* has been previously shown to produce sulphur compounds only in fermented meat (56, 57), in this study was correlated with specific sulphur compounds in cheese. This data would suggest that some smear bacteria though present at relatively low abundance in cheese are likely contributors to the release of FFAs and to their degradation, due to esterase activity and hence contribute to the aroma and flavour in the final cheese product (58, 59).

In the study reported here, whole-metagenome shotgun sequencing was employed as a novel method for the analysis of a fermented product with a complex microbiota. Metagenomic analysis was an efficient tool to understand the variations of the microbial population of the cheese surface over time and the related metabolic potential. Moreover, the association between the volatile compounds and the species represents a novel system to study the flavour development in cheese. In conclusion, the approach used in this study enabled us to determine the microbial succession during ripening, and also to begin to unravel the contributions of the various components of the surface microbiota when present within a complex microbial environment. The method proposed in this study can be adopted in industry to control

the microbiota of fermented food, driving to the production of food products with specific flavour characteristics.

SUPPLEMENTAL MATERIAL

Table S1: Relative abundance (%) of the microbial species within D4 and S5 mix. Data are the mean of 3 replicates. Species highlighted in bold were stated as present by the culture provider.

Species in the smear-culture mixes (%)	D4 mix	S5 mix
Brevibacterium casei	**61.09**	-[nd]
Brevibacterium linens	**5.26**	**3.53**
Glutamicibacter arilaitensis	**-**	**64.03**
Staphylococcus xylosus	**-**	**0.57**
Cyberlindnera jadinii	**14.84**	-
Debaryomyces hansenii	**1.88**	**14.66**
Geotrichum candidum	**-**	**12.12**
Brevibacterium sp. VCM10	12.82	
Brevibacterium siliguriense	1.41	-
Brevibacterium epidermidis	1.1	-
Brevibacterium sandarakinum	0.59	-
Arthrobacter sp. NIO-1057	-	1.27
Debaryomyces fabryi	-	1.07
Arthrobacter sp. W1	-	0.38
Glutamicibacter mysorens	-	0.24
Arthrobacter sp. EpRS66	-	0.16
Paeniglutamicibacter antarcticus	-	0.14
Others	1	1.81

Table S2: Relative abundance of the microbial species on the cheese surface of control, D4 and S5 at day 0, 18, 24 and 30. Data are the mean of 3 replicates.

	Control				D4				S5			
Species	**0**	**18**	**24**	**30**	**0***	**18**	**24**	**30**	**0***	**18**	**24**	**30**
Lactococcus lactis	75.85	57.58	57.76	65.99	74.35	18.7	21.72	6.18	74.74	13.59	11.02	13.94
Streptococcus thermophilus	19.65	36.93	36.53	28.21	19.4	6.03	11.66	4.98	20.04	5.79	5.91	5.82
Glutamicibacter arilaitensis	-[nd]	-	0.24	-	-	30.9	43.27	73.75	-	0.26	0.44	0.42
Debaryomyces hansenii	-	-	0.07	0.45	-	34.12	13.29	4.14	-	21.2	9.57	14.09
Geotrichum candidum	-	-	0.08	0.28	-	-	0.31	1.11	-	37.54	17.6	26.44
Brevibacterium linens	-	-	-	-	-	-	0.17	0.26	-	8.84	37.05	22.84
Staphylococcus xylosus	-	-	-	-	-	0.11	-	-	-	9.08	13.36	10.83
Lactobacillus helveticus	2.12	3.1	2.78	2.72	2.38	1.39	1.41	0.34	2.19	0.86	0.48	0.46
Acinetobacter baumannii	0.82	0.12	0.17	0.37	0.63	-	-	-	0.95	-	-	0.15
Streptococcus pneumoniae	0.56	0.79	0.74	0.66	0.95	-	0.29	0.06	0.74	0.12	0.06	-
Streptococcus salivarius	0.5	0.93	0.93	0.71	0.52	-	0.3	-	0.5	0.11	0.06	0.05
Arthrobacter sp. NIO-1057	-	-	-	-	-	0.57	0.72	1.29	-	-	-	-
Staphylococcus equorum	-	-	-	-	-	1.32	0.43	0.45	-	-	-	-
Staphylococcus saprophyticus	-	-	-	-	-	2.69	0.93	1.06	-	-	-	-
Penicillium camemberti	-	-	-	-	-	0.37	0.4	0.63	-	-	-	-
Corynebacterium variabile	-	-	-	-	-	-	2.04	2.08	-	-	-	-
Debaryomyces fabryi	-	-	-	-	-	1.47	0.56	0.13	-	1.64	0.71	1.09
Psychrobacter sp. P11F6	-	-	-	-	-	-	-	-	-	-	0.4	0.5
Psychrobacter glacincola	-	-	-	-	-	-	-	-	-	-	0.83	1.13
Psychrobacter sp. JCM 18903	-	-	-	-	-	-	-	-	-	-	0.52	0.65
Stenotrophomonas maltophilia	-	-	0.2	0.37	-	-	-	-	-	0.17	-	0.18
Brevibacterium sandarakinum	-	-	-	-	-	-	-	-	-	0.2	0.94	0.58
Anaplasma phagocytophilum	0.22	-	-	-	0.74	-	-	-	0.45	-	-	-
Other	0.27	0.54	0.49	0.24	1.04	2.33	2.5	3.56	0.39	0.59	1.03	0.86

Table S3: Reference genomes used to construct PanPhlAn pangenome databases.

Pangenome Database	Reference Strain (RefSeq Assembly Accession)
Brevibacterium linens	GCF_000167575
	GCF_000807915
	GCF_001606005
	GCF_001729525
Glutamicibacter arilaitensis	GCF_000197735
	GCF_000238915
	GCF_001302565
	GCF_002189495
Lactococcus lactis	GCF_000014545
	GCF_000025045
	GCF_000143205
	GCF_000312685
	GCF_000348965
	GCF_000442845
	GCF_000447825
	GCF_000447845
	GCF_000447885
	GCF_000447985
	GCF_000468955
	GCF_000479375
	GCF_000488975
	GCF_000493355
	GCF_000534815
	GCF_000615405
	GCF_000731635
	GCF_000761115
	GCF_000786755
Streptococcus thermophilus	GCF_000011825
	GCF_000011845
	GCF_000014485
	GCF_000182875
	GCF_000253395
	GCF_000262675
	GCF_000284675
	GCF_000335495
	GCF_000335515
	GCF_000434755
	GCF_000500565
	GCF_000521265
	GCF_000521285

	GCF_000521305
	GCF_000521325
	GCF_000572065
	GCF_000572095
	GCF_000698885
	GCF_000836675
	GCF_000971665
	GCF_001068405
	GCF_001071365
	GCF_001073445
Staphylococcus xylosus	GCF_000338275
	GCF_000467225
	GCF_000706685
	GCF_000709415
	GCF_000815285
	GCF_000953575
	GCF_001476985
	GCF_001747725
	GCF_001747735
	GCF_001747745
	GCF_001748025
	GCF_001748045
	GCF_002078255
	GCF_900098615

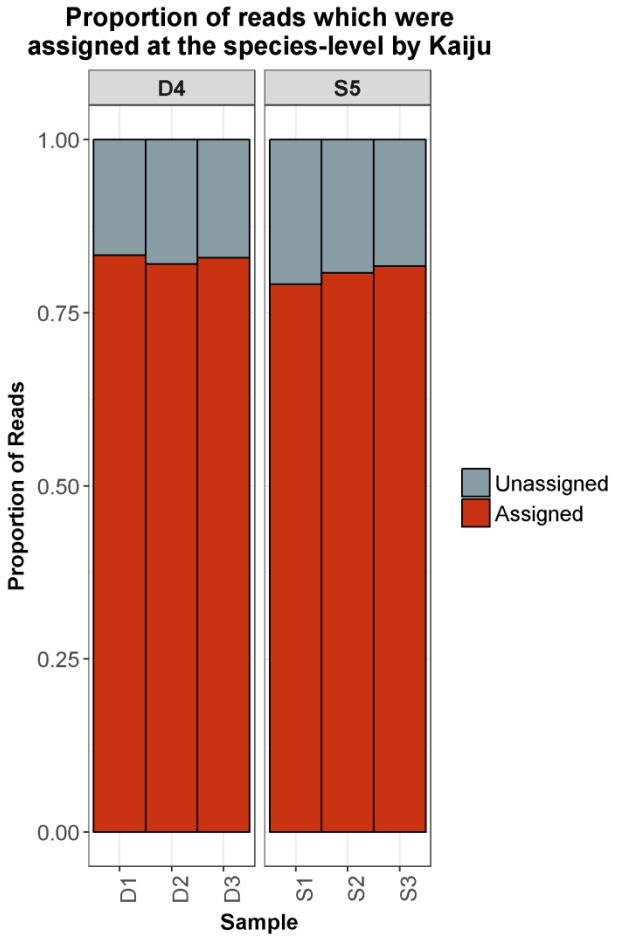


Figure S1: Proportions of reads assigned to the species-level by Kaiju

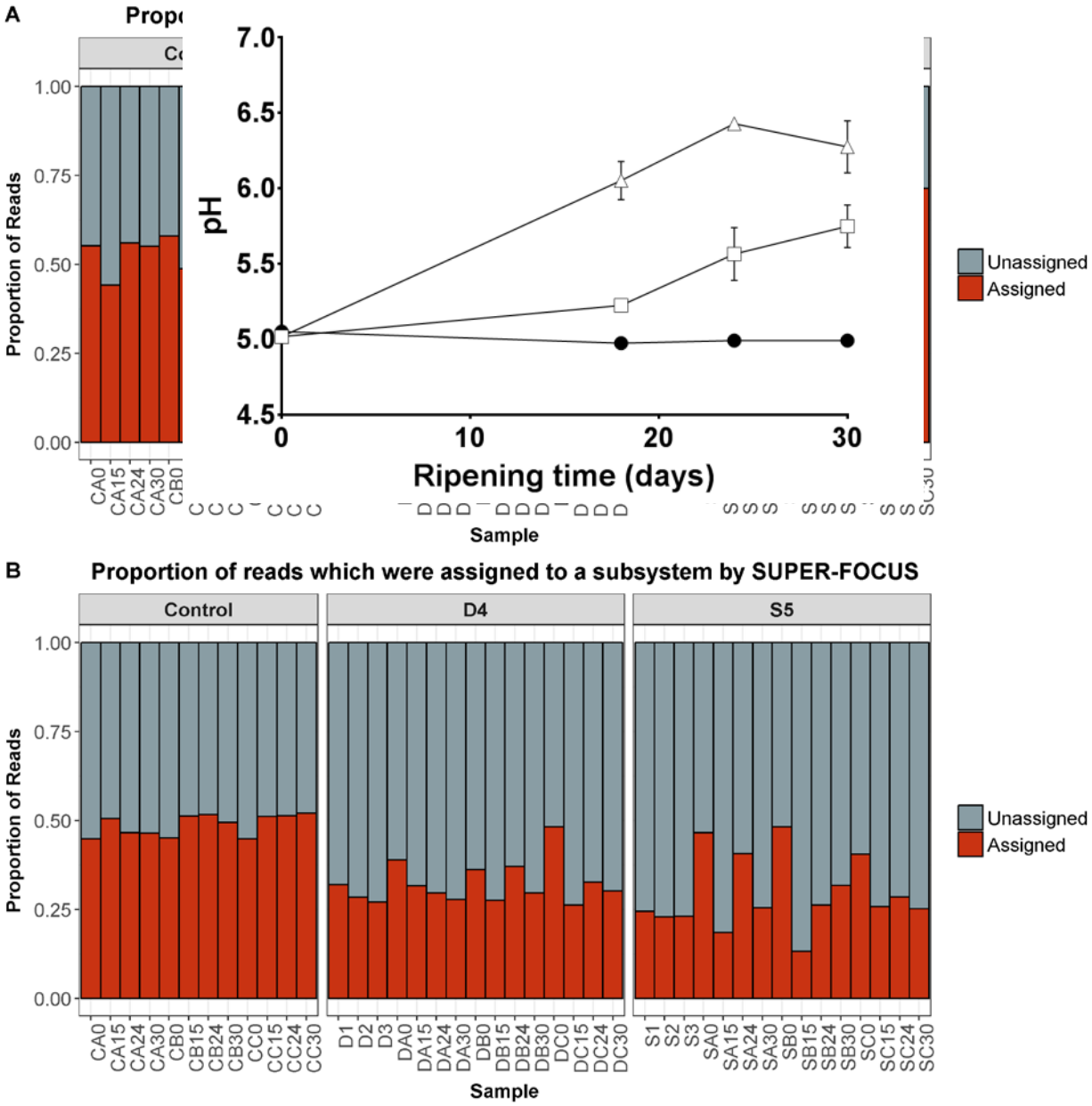


Figure S2: Proportions of assigned reads with Kaiju (A) and SUPER-FOCUS (B) for samples cheese surface samples.

Figure S3. **Changes in the pH values of the surfaces of the control (circles), D4 (squares), and S5 (triangles) cheeses. Data show the means and standard deviations of results from three replicate trials.**

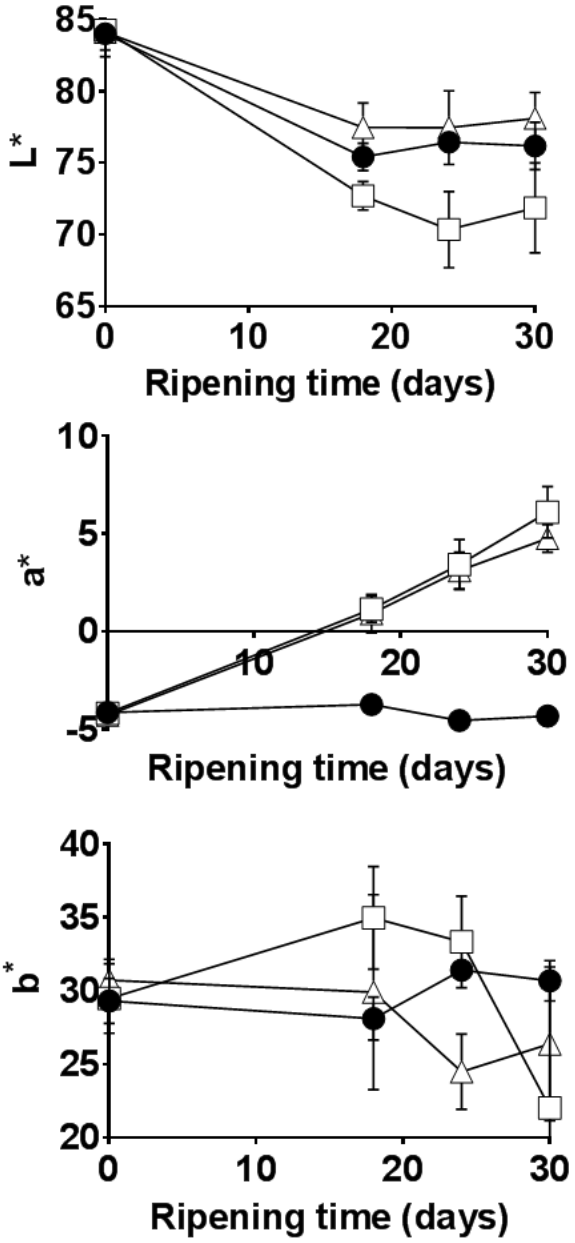


Figure S4: Color development on the surfaces of the control (circles), D4 (squares), and S5 (triangles) cheeses. Data show the means and standard deviations of results from three replicate trials.

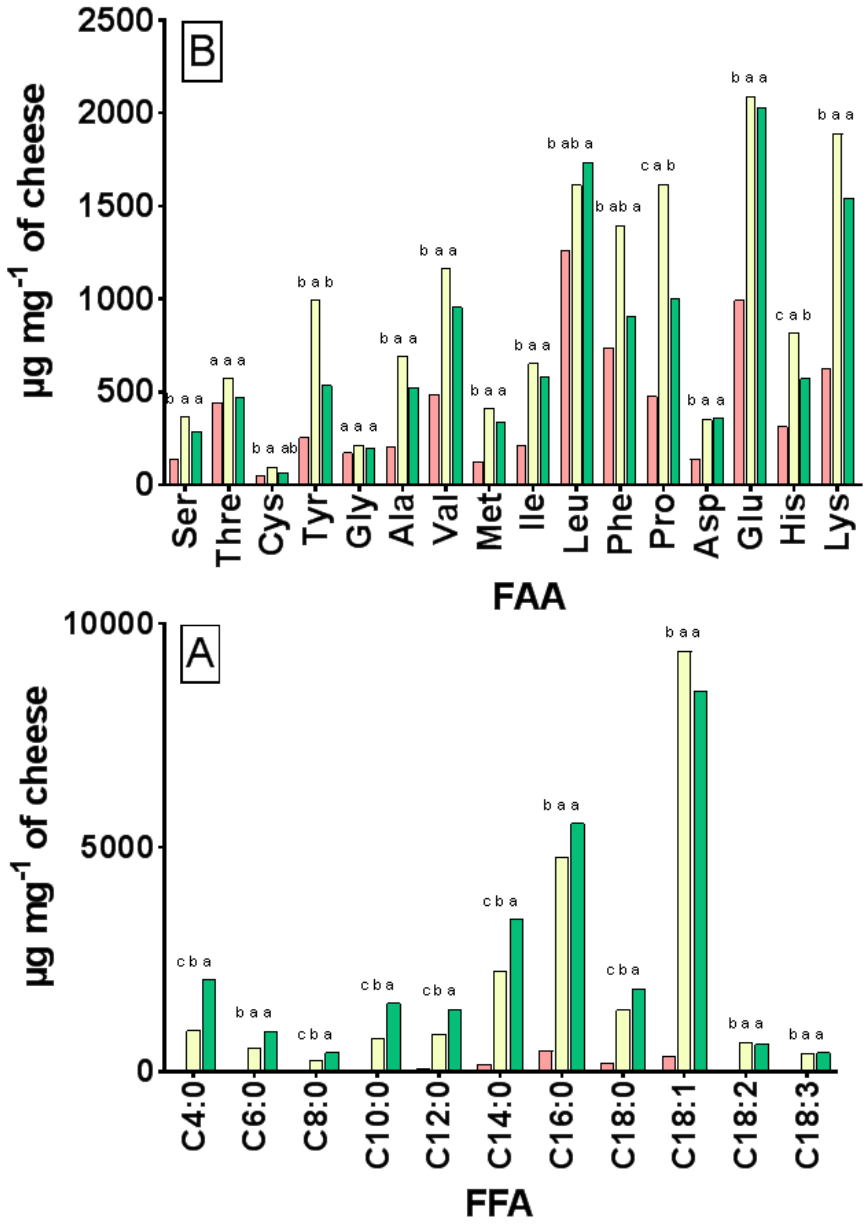


Figure S5: Free amino acid (A) and free fatty acid (B) concentrations (micrograms per milligram) on the surfaces of the control (red), D4 (green), and S5 (yellow) cheeses at day 30. Data show the means of results from three replicate trials. The significant differences ($P < 0.05$) are indicated with a, b, and c.

References

1. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.
2. **Dugat-Bony E, Straub C, Teissandier A, Onésime D, Loux V, Monnet C, Irlinger F, Landaud S, Leclercq-Perlat M-N, Bento P.** 2015. Overview of a surface-ripened cheese community functioning by meta-omics analyses. PLoS One **10:**e0124360.
3. **Wolfe BE, Button JE, Santarelli M, Dutton RJ.** 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell **158:**422-433.
4. **Wolfe Benjamin E, Dutton Rachel J.** 2015. Fermented Foods as Experimentally Tractable Microbial Ecosystems. Cell **161:**49-55.
5. **Montel M-C, Buchin S, Mallet A, Delbes-Paus C, Vuitton DA, Desmasures N, Berthier F.** 2014. Traditional cheeses: Rich and diverse microbiota with associated benefits. International Journal of Food Microbiology **177:**136-154.
6. **Mounier J, Coton M, Irlinger F, Landaud S, Bonnarme P.** 2017. Chapter 38 - Smear-Ripened Cheeses, p 955-996, Cheese (Fourth edition) doi:https://doi.org/10.1016/B978-0-12-417012-4.00038-7. Academic Press, San Diego.
7. **Desmasures N, Bora N, Ward AC.** 2015. Smear Ripened Cheeses, p 1-18. *In* Bora N, Dodd C, Desmasures N (ed), Diversity, Dynamics and Functional

Role of Actinomycetes on European Smear Ripened Cheeses doi:10.1007/978-3-319-10464-5_1. Springer International Publishing, Cham.

8. **Quigley L, O'Sullivan O, Beresford TP, Ross RP, Fitzgerald GF, Cotter PD.** 2012. High-throughput sequencing for detection of subpopulations of bacteria not previously associated with artisanal cheeses. Applied and Environmental Microbiology **78:**5717-5723.

9. **Delcenserie V, Taminiau B, Delhalle L, Nezer C, Doyen P, Crevecoeur S, Roussey D, Korsak N, Daube G.** 2014. Microbiota characterization of a Belgian protected designation of origin cheese, Herve cheese, using metagenomic analysis. Journal of Dairy Science **97:**6046-6056.

10. **Bokulich NA, Mills DA.** 2013. Facility-specific “house” microbiome drives microbial landscapes of artisan cheesemaking plants. Applied and Environmental Microbiology **79:**5214-5223.

11. **Bertuzzi AS, Kilcawley KN, Sheehan JJ, O'Sullivan MG, Kennedy D, McSweeney PLH, Rea MC.** 2017. Use of smear bacteria and yeasts to modify flavour and appearance of Cheddar cheese. International Dairy Journal **72:**44-54.

12. **Roy J.** 2007. SAS for Mixed Models, Second Edition. R. C.Littell, G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. Journal of Biopharmaceutical Statistics **17:**363-365.

13. **Clooney AG, Fouhy F, Sleator RD, O’Driscoll A, Stanton C, Cotter PD, Claesson MJ.** 2016. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PLoS One **11:**e0148028.

14. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

15. **Menzel P, Ng KL, Krogh A.** 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications **7:**11257.

16. **Pruitt KD, Tatusova T, Maglott DR.** 2006. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research **35:**D61-D65.

17. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

18. **Silva GGZ, Green KT, Dutilh BE, Edwards RA.** 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics **32:**354-361.

19. **Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R.** 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research **33:**5691-5702.

20. **McDermott A, Visentin G, De Marchi M, Berry DP, Fenelon MA, O'Connor PM, Kenny OA, McParland S.** 2016. Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. Journal of Dairy Science **99:**3171-3182.

21. **Catrienus DJ, T. BH.** 1990. Determination of free fatty acids in milk and cheese procedures for extraction, clean up, and capillary gas chromatographic analysis. Journal of High Resolution Chromatography **13:**94-98.

22. **Mannion DT, Furey A, Kilcawley KN.** 2016. Comparison and validation of 2 analytical methods for the determination of free fatty acids in dairy products by gas chromatography with flame ionization detection. Journal of Dairy Science **99:**5047-5063.

23. **Team RC.** 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.

24. **Ferreira AD, Viljoen BC.** 2003. Yeasts as adjunct starters in matured Cheddar cheese. International Journal of Food Microbiology **86:**131-140.

25. **Mounier J.** 2015. Microbial Interactions in Smear-Ripened Cheeses, p 155-166. *In* Bora N, Dodd C, Desmasures N (ed), Diversity, Dynamics and Functional Role of Actinomycetes on European Smear Ripened Cheeses doi:10.1007/978-3-319-10464-5_6. Springer International Publishing, Cham.

26. **Zikánová B, Kuthan M, Řičicová M, Forstová J, Palková Z.** 2002. Amino acids control ammonia pulses in yeast colonies. Biochemical and Biophysical Research Communications **294:**962-967.

27. **Corsetti A, Rossi J, Gobbetti M.** 2001. Interactions between yeasts and bacteria in the smear surface-ripened cheeses. International Journal of Food Microbiology **69:**1-10.

28. **Hickey CD, Fallico V, Wilkinson MG, Sheehan JJ.** 2018. Redefining the effect of salt on thermophilic starter cell viability, culturability and metabolic activity in cheese. Food Microbiology **69:**219-231.

29. **Monnet C, Landaud S, Bonnarme P, Swennen D.** 2015. Growth and adaptation of microorganisms on the cheese surface. FEMS Microbiology Letters **362:**1-9.

30. **Krubasik P, Sandmann G.** 2000. A carotenogenic gene cluster from Brevibacterium linens with novel lycopene cyclase genes involved in the synthesis of aromatic carotenoids. Molecular and General Genetics MGG **263:**423-432.

31. **Mounier J, Irlinger F, Leclercq-Perlat M-N, Sarthou A-S, Spinnler H-E, Fitzgerald GF, Cogan TM.** 2006. Growth and colour development of some surface ripening bacteria with Debaryomyces hansenii on aseptic cheese curd. Journal of Dairy Research **73:**441-448.

32. **Feurer C, Vallaeys T, Corrieu G, Irlinger F.** 2004. Does Smearing Inoculum Reflect the Bacterial Composition of the Smear at the End of the Ripening of a French Soft, Red-Smear Cheese. Journal of Dairy Science **87:**3189-3197.

33. **Goerges S, Mounier J, Rea MC, Gelsomino R, Heise V, Beduhn R, Cogan TM, Vancanneyt M, Scherer S.** 2008. Commercial ripening starter microorganisms inoculated into cheese milk do not successfully establish themselves in the resident microbial ripening consortia of a South german red smear cheese. Appl Environ Microbiol **74:**2210-2217.

34. **Rea MC, Görges S, Gelsomino R, Brennan NM, Mounier J, Vancanneyt M, Scherer S, Swings J, Cogan TM.** 2007. Stability of the Biodiversity of the Surface Consortia of Gubbeen, a Red-Smear Cheese. Journal of Dairy Science **90:**2200-2210.

35. **Larpin-Laborde S, Imran M, Bonaiti C, Bora N, Gelsomino R, Goerges S, Irlinger F, Goodfellow M, Ward AC, Vancanneyt M, Swings J, Scherer S, Gueguen M, Desmasures N.** 2011. Surface microbial consortia from Livarot, a French smear-ripened cheese. Can J Microbiol **57:**651-660.

36. **Gori K, Ryssel M, Arneborg N, Jespersen L.** 2013. Isolation and Identification of the Microbiota of Danish Farmhouse and Industrially Produced Surface-Ripened Cheeses. Microbial Ecology **65:**602-615.

37. **Cholet O, Henaut A, Casaregola S, Bonnarme P.** 2007. Gene expression and biochemical analysis of cheese-ripening yeasts: focus on catabolism of L-methionine, lactate, and lactose. Appl Environ Microbiol **73:**2561-2570.

38. **Yvon M, Rijnen L.** 2001. Cheese flavour formation by amino acid catabolism. International Dairy Journal **11:**185-201.

39. **Collins YF, McSweeney PLH, Wilkinson MG.** 2003. Lipolysis and free fatty acid catabolism in cheese: a review of current knowledge. International Dairy Journal **13:**841-866.

40. **Leclercq-Perlat MN, Corrieu G, Spinnler HE.** 2004. Comparison of Volatile Compounds Produced in Model Cheese Medium Deacidified by Debaryomyces hansenii or Kluyveromyces marxianus. Journal of Dairy Science **87:**1545-1550.

41. **Klaus G, Marie SL, Agerlin PM, Lene J, Nils A.** 2012. Debaryomyces hansenii strains differ in their production of flavor compounds in a cheese-surface model. MicrobiologyOpen **1:**161-168.

42. **Arfi K, Spinnler H, Tache R, Bonnarme P.** 2002. Production of volatile compounds by cheese-ripening yeasts: requirement for a methanethiol donor

for S-methyl thioacetate synthesis by Kluyveromyces lactis. Applied Microbiology and Biotechnology **58:**503-510.

43. **Martin N, Berger C, Le Du C, Spinnler HE.** 2001. Aroma Compound Production in Cheese Curd by Coculturing with Selected Yeast and Bacteria. Journal of Dairy Science **84:**2125-2135.

44. **Deetae P, Bonnarme P, Spinnler HE, Helinck S.** 2007. Production of volatile aroma compounds by bacterial strains isolated from different surface-ripened French cheeses. Applied Microbiology and Biotechnology **76:**1161-1171.

45. **Monnet C, Loux V, Gibrat J-F, Spinnler E, Barbe V, Vacherie B, Gavory F, Gourbeyre E, Siguier P, Chandler M, Elleuch R, Irlinger F, Vallaeys T.** 2010. The Arthrobacter arilaitensis Re117 Genome Sequence Reveals Its Genetic Adaptation to the Surface of Cheese. PLoS One **5:**e15489.

46. **Mounier J, Monnet C, Vallaeys T, Arditi R, Sarthou AS, Helias A, Irlinger F.** 2008. Microbial interactions within a cheese microbial community. Appl Environ Microbiol **74:**172-181.

47. **Boutrou R, Guéguen M.** 2005. Interests in Geotrichum candidum for cheese technology. International Journal of Food Microbiology **102:**1-20.

48. **Jollivet N, Chataud J, Vayssier Y, Bensoussan M, Belin J-M.** 1994. Production of volatile compounds in model milk and cheese media by eight strains of Geotrichum candidum Link. Journal of Dairy Research **61:**241-248.

49. **Brennan NM, Ward AC, Beresford TP, Fox PF, Goodfellow M, Cogan TM.** 2002. Biodiversity of the bacterial flora on the surface of a smear cheese. Appl Environ Microbiol **68:**820-830.

50. **Mounier J, Gelsomino R, Goerges S, Vancanneyt M, Vandemeulebroecke K, Hoste B, Scherer S, Swings J, Fitzgerald GF, Cogan TM.** 2005. Surface microflora of four smear-ripened cheeses. Appl Environ Microbiol **71:**6489-6500.

51. **Lecocq J, Gueguen M.** 1994. Effects of pH and Sodium Chloride on the Interactions Between Geotrichum candidum and Brevibacterium linens. Journal of Dairy Science **77:**2890-2899.

52. **Rattray FP, Fox PF.** 1999. Aspects of Enzymology and Biochemical Properties of Brevibacterium linens Relevant to Cheese Ripening: A Review1. Journal of Dairy Science **82:**891-909.

53. **Jollivet N, Bézenger M-C, Vayssier Y, Belin J-M.** 1992. Production of volatile compounds in liquid cultures by six strains of coryneform bacteria. Applied Microbiology and Biotechnology **36:**790-794.

54. **Irlinger F, Morvan A, El Solh N, Bergere JL.** 1997. Taxonomic Characterization of Coagulase-Negative Staphylococci in Ripening Flora from Traditional French Cheeses. Systematic and Applied Microbiology **20:**319-328.

55. **Mounier J, Goerges S, Gelsomino R, Vancanneyt M, Vandemeulebroecke K, Hoste B, Brennan NM, Scherer S, Swings J, Fitzgerald GF, Cogan TM.** 2006. Sources of the adventitious microflora of a smear-ripened cheese. J Appl Microbiol **101:**668-681.

56. **Stahnke LH.** 1999. Volatiles Produced by Staphylococcus xylosus and Staphylococcus carnosus during Growth in Sausage Minces Part I. Collection and Identification. LWT - Food Science and Technology **32:**357-364.

57. **Tjener K, Stahnke LH, Andersen L, Martinussen J.** 2004. The pH-unrelated influence of salt, temperature and manganese on aroma formation by Staphylococcus xylosus and Staphylococcus carnosus in a fermented meat model system. International Journal of Food Microbiology **97:**31-42.

58. **Curtin ÁC, Gobbetti M, McSweeney PLH.** 2002. Peptidolytic, esterolytic and amino acid catabolic activities of selected bacterial strains from the surface of smear cheese. International Journal of Food Microbiology **76:**231-240.

59. **Casaburi A, Villani F, Toldrá F, Sanz Y.** 2006. Protease and esterase activity of staphylococci. International Journal of Food Microbiology **112:**223-229.

Chapter 5

Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks

Figures updated since publication in *Applied and Environmental Microbiology*

(doi: https://doi.org/10.1128/AEM.01144-17)

Authors: Aaron M. Walsh, Fiona Crispie, Kareem Daari, Orla O'Sullivan, Jennifer C. Martin, Cornelius T. Arthur, Marcus J. Claesson, Karen P. Scott, and Paul D. Cotter

Contributions:

- **Candidate** performed sequencing library preparations, bioinformatics analysis, and statistical analysis
- **KD**, **JCM**, and **CTA** performed culture work and DNA extractions
- **OOS** provided guidance for bioinformatic analysis
- **FC**, **MJC**, **KPS**, and **PDC** supervised the study

Abstract

The rapid detection of pathogenic strains in food products is essential for the prevention of disease outbreaks. It has already been demonstrated that whole metagenome shotgun sequencing can be used to detect pathogens in food but, until recently, strain-level detection of pathogens has relied on whole metagenome assembly, which is a computationally demanding process. Here, we demonstrate that three short read alignment-based methods, MetaMLST, PanPhlAn, and StrainPhlAn, can accurately, and rapidly, identify pathogenic strains in spinach metagenomes which were intentionally spiked with Shiga toxin-producing *Escherichia coli* in a previous study. Subsequently, we employ the methods, in combination with other metagenomics approaches, to assess the safety of nunu, a traditional Ghanaian fermented milk product which is produced by the spontaneous fermentation of raw cow milk. We show that nunu samples are frequently contaminated with bacteria associated with the bovine gut, and worryingly, we detect putatively pathogenic *E. coli* and *Klebsiella pneumoniae* strains in a subset of nunu samples. Ultimately, our work establishes that short read alignment-based bioinformatics approaches are suitable food safety tools, and we describe a real-life example of their utilisation.

Introduction

In recent years, high-throughput sequencing (HTS) has become an important tool in food microbiology (1). HTS enables in-depth characterisation of food-related microbial isolates, *via* whole genome sequencing (WGS), and it facilitates culture-independent analysis of mixed microbial communities in foods, *via* metagenomic sequencing.

WGS has provided invaluable insights into the genetics of starter cultures (2, 3), and it is routinely used in epidemiology to identify outbreak-associated foodborne pathogens isolated from clinical samples, by comparing the single nucleotide polymorphism (SNP) profiles of outbreak strain genomes versus non-outbreak strain genomes (4-6). Metagenomic sequencing enables the elucidation of the roles of microorganisms during food production (7-9), and it can be used to track microorganisms of interest through the food production chain, as illustrated by Yang *et al.* (10), who used whole metagenome shotgun sequencing to track pathogenic species in the beef production chain. Indeed, metagenomic sequencing can be used to detect pathogens in foods to monitor outbreaks of foodborne illnesses (11), but few studies have done so, because of the limited taxonomic resolution achievable using these methods. Typically, 16S rRNA gene sequencing provides genus-level taxonomic resolution (12), and although sub-genus-level classification is achievable using species-classifiers (13) or oligotyping (14, 15), these methods cannot accurately discriminate between strains. Similarly, metagenome sequence classification tools usually provide species-level resolution (16). However, strain-level resolution is necessary for the accurate identification of pathogens in food products (17). Leonard *et al.* successfully achieved strain-level resolution of Shiga toxin producing *Escherichia coli* strains in spinach samples using metagenome

shotgun sequencing (18). However, the bioinformatics methods used in that study were based on metagenome assembly, which is a computationally demanding process (19, 20), and thus alternative strain-level identification methods are needed.

Since 2016, several short read alignment based software applications, including MetaMLST (20), StrainPhlAn (21), and PanPhlAn (19), have been released that can achieve strain-level characterisation of microorganisms from metagenome shotgun sequencing data. All three applications are considerably faster than metagenome assembly based methods. To date, these programs have not been employed to detect pathogens in food products, but there is strong evidence to suggest that they have considerable potential for this purpose: MetaMLST accurately predicted that the strain responsible for the 2011 German *E. coli* outbreak belonged to *E. coli* ST678 (20), and similarly, PanPhlAn accurately predicted that the strain was a Shiga toxin producer (19), based on the analysis of the gut metagenomes of infected patients (22). StrainPhlAn has so far not been used for epidemiological purposes, but a recent study demonstrated that it can be used to predict the phylogenetic relatedness of bacterial strains from different samples (21).

MetaMLST aligns sequencing reads against a housekeeping gene database to identify sequence types present in metagenomic samples based on multilocus sequence typing (MLST). The MetaMLST database contains all currently known sequence types, but it can be updated as required to include newly identified sequence types. MetaMLST does not require any prior knowledge of the microbial composition of sample and it can simultaneously detect different species' sequence types. PanPhlAn aligns sequencing reads against a species pangenome database, constructed from reference genomes, to functionally characterise strains present in metagenomic samples. PanPhlAn allows the user to generate customisable

pangenome databases for any species. StrainPhlAn extracts species specific marker genes from sequencing reads and it aligns the markers against reference genomes to identify the strains present in metagenomic samples. StrainPhlAn requires output from MetaPhlAn2, and both programs use the same database.

In this study, we describe the characterisation of nunu, a traditional Ghanaian fermented milk product (FMP), at the genus, species, and strain-levels, using a combination of 16S rRNA gene sequencing and whole metagenome shotgun sequencing. Nunu is produced by the spontaneous fermentation of raw cow milk in calabashes or plastic or metal containers under ambient conditions, and it is usually consumed after 24-36 hours (23). At present, little is known about nunu's microbiology, relative to other FMPs, like kefir or yoghurt (24). Previously, a number of potentially pathogenic bacteria, including *Enterobacter*, *Escherichia* and *Klebsiella*, were detected in nunu by culture based methods (25). Here, we carry out the first culture-independent analysis of a number of nunu samples. In addition to detecting the presence of a variety of lactic acid bacteria (LAB) typical of fermented dairy products, MetaMLST, PanPhlAn and StrainPhlAn all indicated the presence of pathogenic *E. coli* and *Klebsiella pneumoniae* in a subset of the samples. We also demonstrate that these tools can accurately predict the presence of pathogenic strains in foods by testing them on food metagenomes which were spiked with Shiga toxin producing *E. coli*. Ultimately, our work establishes that short read alignment based methods can be used for the detection of pathogens in foods.

Materials and Methods

Sampling

Five nunu samples were collected from producers with hygiene practice training, and another five samples were collected from producers without hygiene practice training. The identity of the samples from trained and untrained individuals was blinded until after sequencing analysis was completed. The samples from the trained group were labelled 1t2am, 1t6am, 1t7am, 1t8am, and 2t2am. The samples from the untrained group labelled 1u6am, 2u2am, 2u3am, 2u6am, and 2u8am. All samples were collected in the morning and placed on ice for transport to the lab. Sample aliquots (4ml) were then mixed with glycerol to a final concentration of 20% and stored at -20°C prior to DNA extraction. DNA was extracted from the samples at the Animal Research Institute, Accra, Ghana and then sent to Scotland to comply with International laws on the import of animal samples (Import Licence form AB117).

Microbiological analysis

Basic microbiology culture analysis was carried out in Ghana. The plate-count technique was used to estimate the total viable bacterial count of the nunu samples on Milk Plate Count Agar (LAB M, UK). Bacterial counts were compared for plates growing aerobically or anaerobically at 30°C for 36-72 h. Anaerobic plates were incubated in airtight canisters containing $C0_2$Gen sachets (Oxoid, UK), which created an anaerobic atmosphere. Following incubation, colonies were counted using an SC6+ electronic colony counter (Stuart Scientific, UK). The presence of specific pathogens in the nunu samples was determined by streaking nunu directly onto selective agar plates to visually assess bacterial growth. The following selective agars were used: Blood agar (Merck, Germany) for *Staphylococcus*; MacConkey agar (Merck, Germany) for Enterobacteria; de Man Rogosa Sharpe agar (MRS)

(Oxoid, UK) for *Lactobacillus* species; and *Salmonella Shigella* agar (Oxoid, UK). Any mixed growth plates were re-purified by streaking onto selected secondary agars. Lactose fermenting colonies identified on MacConkey agar were sub-cultured onto Eosin Methylene Blue Agar (EMBA) (Scharlau Chemie, Spain) to isolate/identify *E. coli.* Additionally, *Staphylococcus* colonies from Blood Agar were sub-cultured onto Mannitol Salt Agar (MSA) (Oxoid, UK) to isolate/identify *Staphylococcus aureus*. The following biochemical tests were used to confirm bacterial identification: the Motility Indole Urea (MIU) test; the catalase test; the Triple Sugar Iron (TSI) test; and the Indole Methyl Red Vorges-Proskeur Citrate (IMViC) tests. Cellular morphology was determined by Gram staining as well as microscopic examination.

DNA extraction and next generation sequencing

Briefly, 1 ml of each thawed sample was diluted in 9 ml of sterile PBS, mixed thoroughly using vortex and centrifuged for 10 min (8,000-10,000 g). The bacterial cell pellets were resuspended in 432 µl sterile dH_2O and 48µl 0.5 M EDTA, mixed thoroughly by a combination of vortex and with a sterile pipette tip and the suspension frozen. The frozen samples were thawed on the bench and refrozen and finally thawed (giving a total of two freeze/thaw cycles) before extracting the DNA using the Promega Wizard genomic DNA extraction kit (Promega, Madison, WI, USA) according to the manufacturer's protocol. The freeze/thaw cycles were carried out to maximise bacterial cell lysis. Following extraction, the DNA pellets were air dried for about 60 minutes and stored sealed under airtight conditions and

transported from the Animal Research Institute, Accra, Ghana to the Rowett Institute, at University of Aberdeen, for further analysis.

DNA extracts were quantified using the Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland). 16S rRNA gene sequencing libraries were prepared from extracted DNA using the 16S Metagenomic Sequencing Library Preparation protocol from Illumina, with minor modifications (26). Samples were sequenced on the Illumina MiSeq in the Teagasc sequencing facility, with a 2 x 250 cycle V2 kit, in accordance with standard Illumina sequencing protocols. Whole-metagenome shotgun libraries were prepared in accordance with the Nextera XT DNA Library Preparation Guide from Illumina (26). Samples were sequenced on the Illumina MiSeq in the Teagasc sequencing facility, with a 2 x 300 cycle V3 kit, in accordance with standard Illumina sequencing protocols.

Bioinformatics

Raw 16S rRNA gene sequencing reads were quality filtered using PRINSEQ (27). Denoising, OTU clustering, and chimera removal were done using USearch (v7-64bit) (28), as described by Doyle *et al.* (29). OTUs were aligned using PyNAST (30). Alpha-diversity and beta-diversity were calculated using Qiime (1.8.0) (31). Taxonomy was assigned using a BLAST search (32) against SILVA SSU 119 database (33).

Raw whole-metagenome shotgun sequencing reads were filtered, on the basis of quality and quantity, and trimmed to 200 bp, with a combination of Picard Tools (https://github.com/broadinstitute/picard) and SAMtools (34). MetaPhlAn2 was used to characterise the microbial composition of samples at the species-level (35).

MetaMLST (20), PanPhlAn (19), and StrainPhlAn (21) were used to characterise the microbial composition of the samples at the strain-level. GraPhlAn (36) was used to construct phylogenetic trees from the StrainPhlAn output. SUPER-FOCUS (37) and HUMAnN2 (38) were used to determine the microbial metabolic potential of samples. IDBA-UD (39) was used for metagenome assembly.

Sequence data have been deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB20873.

Statistical analysis

Statistical analysis was done in R-3.2.2 (40). The Kruskal-Wallis test was done using the compareGroups package, and the resulting p-values were for multiple comparisons. PCoA analysis of 16S rRNA gene sequencing data was done using the phyloseq package (41). Multidimensional scaling (MDS) was done using the vegan package. Data visualisation was done using the ggplot2 package.

Results

16S rRNA gene sequencing of nunu samples

Nunu samples were collected from producers with hygiene practice training (n=5) and producers without hygiene practice training (n=5), respectively. 16S rRNA gene sequencing analysis revealed that there were no significant differences in the alpha-diversity of nunu samples from trained or untrained producers (Figure S1a), although there was a clear separation in the beta-diversity of the two groups (Figure S1b).

The 16S rRNA data was also analysed to determine bacterial composition (Figure 1a). At the family level, all of the samples were dominated by Lactobacillales, and at the genus-level, most samples were dominated by *Streptococcus*, although the sample 1t2am was dominated by *Lactococcus. Enterococcus* was detected in 4/10 samples (1 trained and 3 untrained) at ≥3% relative abundance, and it was highest in the sample 2u6am, where it was present at 19% relative abundance. In addition, *Staphylococcus* was detected in all 10 samples, although its abundance was ≤1% in each case. The detection of staphylococci was consistent with a corresponding culture-dependent analysis of the samples (supplemental material). Importantly, Enterobacteriales were also prevalent. *Enterobacter* was detected in 9/10 samples (4 samples from trained producers and 5 from untrained producers) at ≥1% relative abundance, and it was highest in the sample 2u8am, where it was present at 23% relative abundance. *Escherichia-Shigella* was detected in 8/10 samples (4 trained and 4 untrained) at ≥1% relative abundance, and it was highest in the sample 1t7am, where it was present at 17% relative abundance; this finding was again consistent with culture-dependent analysis of the samples (supplemental material).

The Kruskal-Wallis test indicated that there were significant differences in the relative abundances of *Macrococcus* (p=0.01), which was higher in samples from trained producers, and *Streptococcus* (p=0.02), which was higher in samples from untrained producers (Figure 1b). No other genera were significantly different.

Species-level compositional analysis of nunu samples as revealed by shotgun sequencing

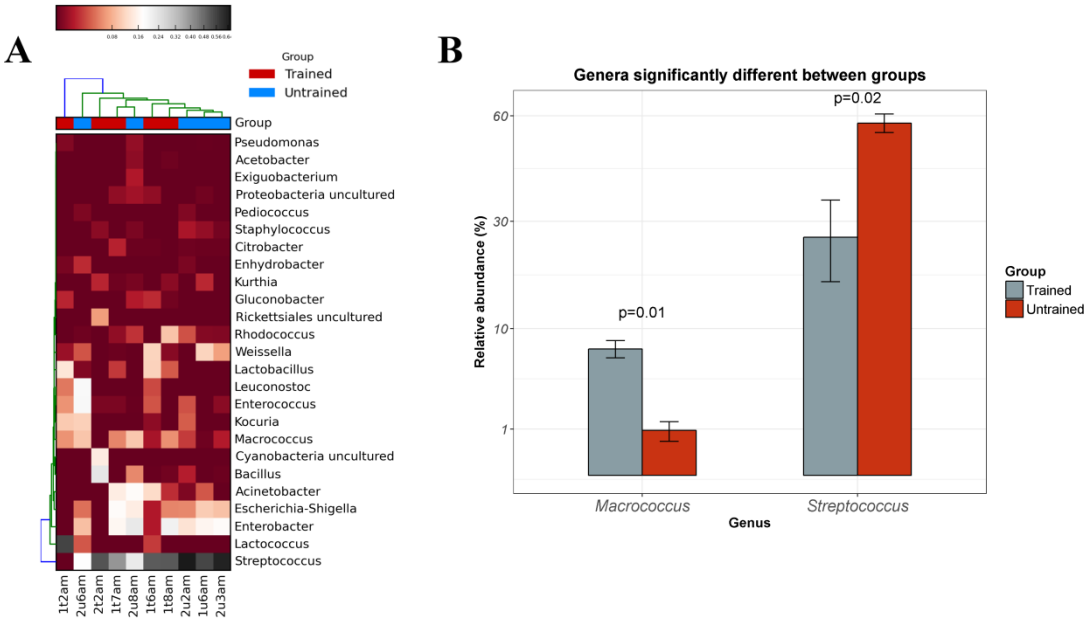


Figure 1: 16S rRNA gene sequencing based analysis of nunu samples. (A) Heat map showing the 25 most abundant bacterial genera across the nunu samples. (B) Bar plot showing genera which were differentially abundant in either group.

MetaPhlAn2-based analysis of shotgun metagenomic data provided results that were generally consistent with those derived from amplicon sequencing. 11 species accounted for >90% of the microbial composition of every sample (Figure 2). At the species-level, most samples were dominated by *Streptococcus infantarius*, although sample 1t2am was dominated by *Lactococcus lactis. Enterococcus faecium* was detected in 4/10 samples (2 trained and 2 untrained) at ≥1% relative abundance, and it was highest in the sample 1t2am, where it was present at 22% relative abundance. High abundances of Enterobacteriales were again apparent. *Enterobacter cloacae* were detected in the sample 1t8am, where it was present at 1% relative abundance. *Escherichia coli* was detected in 2/10 samples (2 trained) at ≥7% relative abundance, and it was highest in 1t7am, where it was present at 13% relative abundance. *Klebsiella pneumoniae* was detected in 7/10 samples (4 trained and 3 untrained) at ≥3% relative abundance, and it was highest in 1t8am, where it was present at 71% relative abundance. In contrast, *Klebsiella* was not detected by amplicon sequencing, and this discrepancy might be due to similarities in the 16S rRNA genes from these genera(42).

The Kruskal-Wallis test indicated that there were significant differences in the relative abundances of *Macrococcus caseolyticus* (p=0.01), which was higher in samples from trained producers, and *Streptococcus infantarius* (p=0.01), which was higher in samples from untrained producers (Figure S2). No other species were significantly different.

Investigation of the functional potential of the nunu microbiota

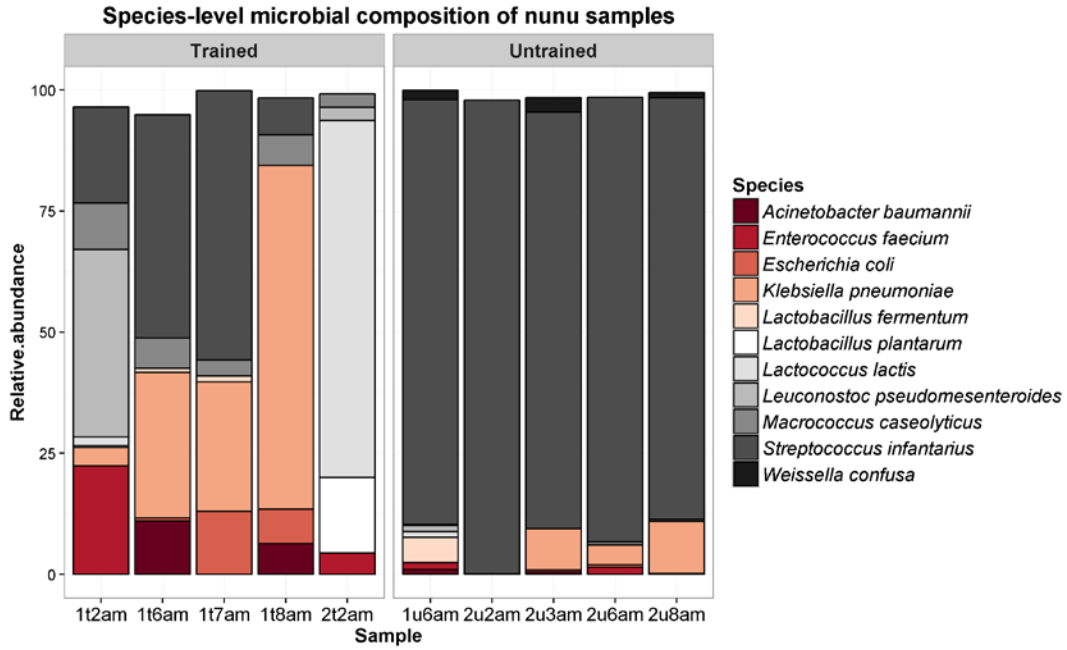


Figure 2: The species-level microbial composition of nunu samples, as determined by MetaPhlAn2.

SUPER-FOCUS was used to provide an overview of the functional potential of the nunu metagenome. As expected, a significant proportion of the metagenome was assigned to housekeeping functions like carbohydrate metabolism, nucleic acid metabolism, and protein metabolism (Figure 3). However, SUPER-FOCUS also detected high levels of functions associated with horizontal gene transfer and virulence in nunu. The level 1 subsystem "Phages, Prophages, Transposable elements" was present at ≥1% average relative abundance in both groups, although it was significantly higher in nunu samples from trained producers ($p=0.047$). Similarly, the level 1 subsystem "Virulence" was present at ≥3.5% average relative abundance in both groups.

HUMAnN2 was used to provide more comprehensive insights into the functional potential of the nunu metagenome. Unsurprisingly, the 25 most abundant genetic pathways were associated with carbohydrate metabolism, nucleic acid metabolism, and protein metabolism (Figure 4a). MDS analysis of all the normalised HUMAnN2 pathway abundances suggested that there were differences in the overall functional potential of the groups (Figure S3), and we detected significant differences in the relative abundances of some individual pathways (Table S1). Notably, we observed that histidine degradation pathways were higher in trained samples ($p=0.047$) (Figure 4c). Furthermore, histidine decarboxylase genes were only detected in trained samples. Several other undesirable genetic pathways were detected in both groups. For example, putrescine biosynthesis pathways and polymyxin resistance genes co-occurred in 7/10 samples (Figure 4c), and these pathways were all attributed to *E. cloacae, E. coli, K. pneumoniae,* or a combination of these three species. We detected several other antibiotic resistance genes, including beta-lactamase genes and methicillin resistance genes, in both groups (Figure S4). In addition, we found HGT-

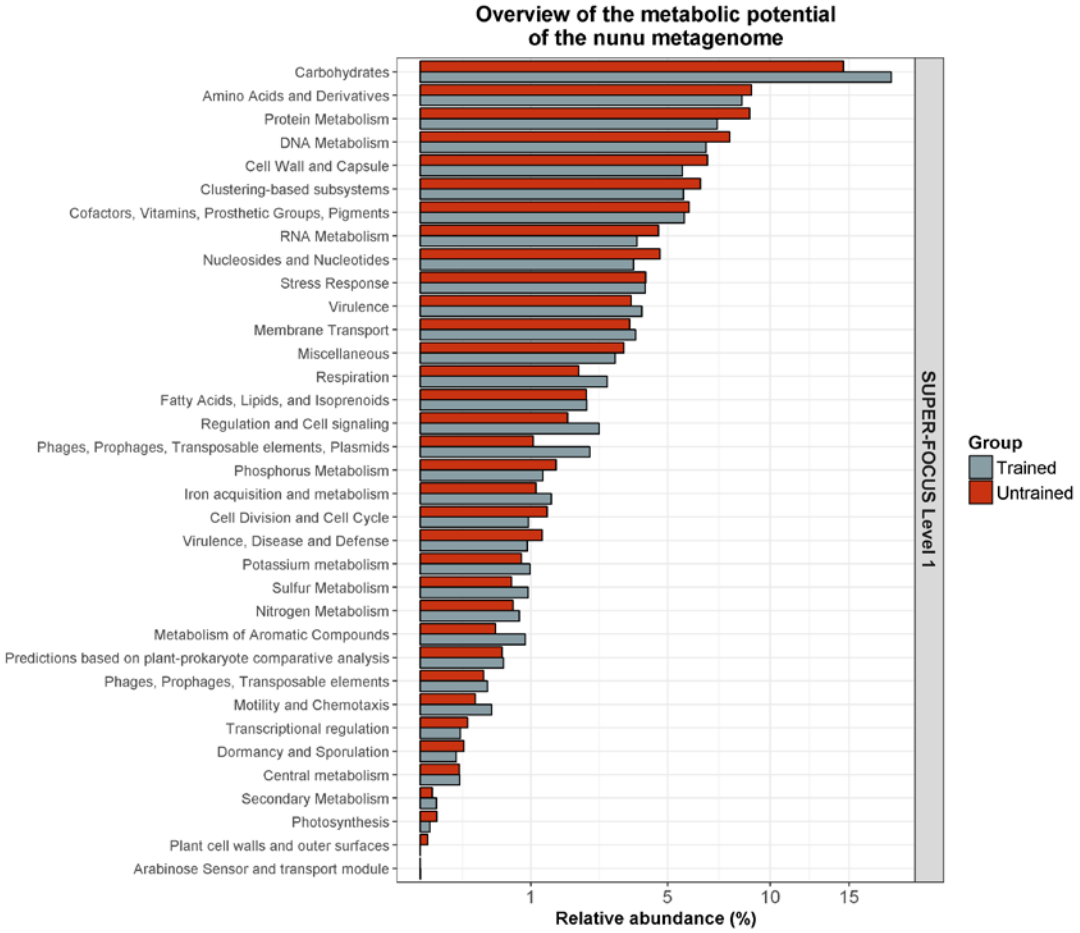


Figure 3: The average abundances of the SUPER-FOCUS Level 1 functions that were detected in nunu samples.

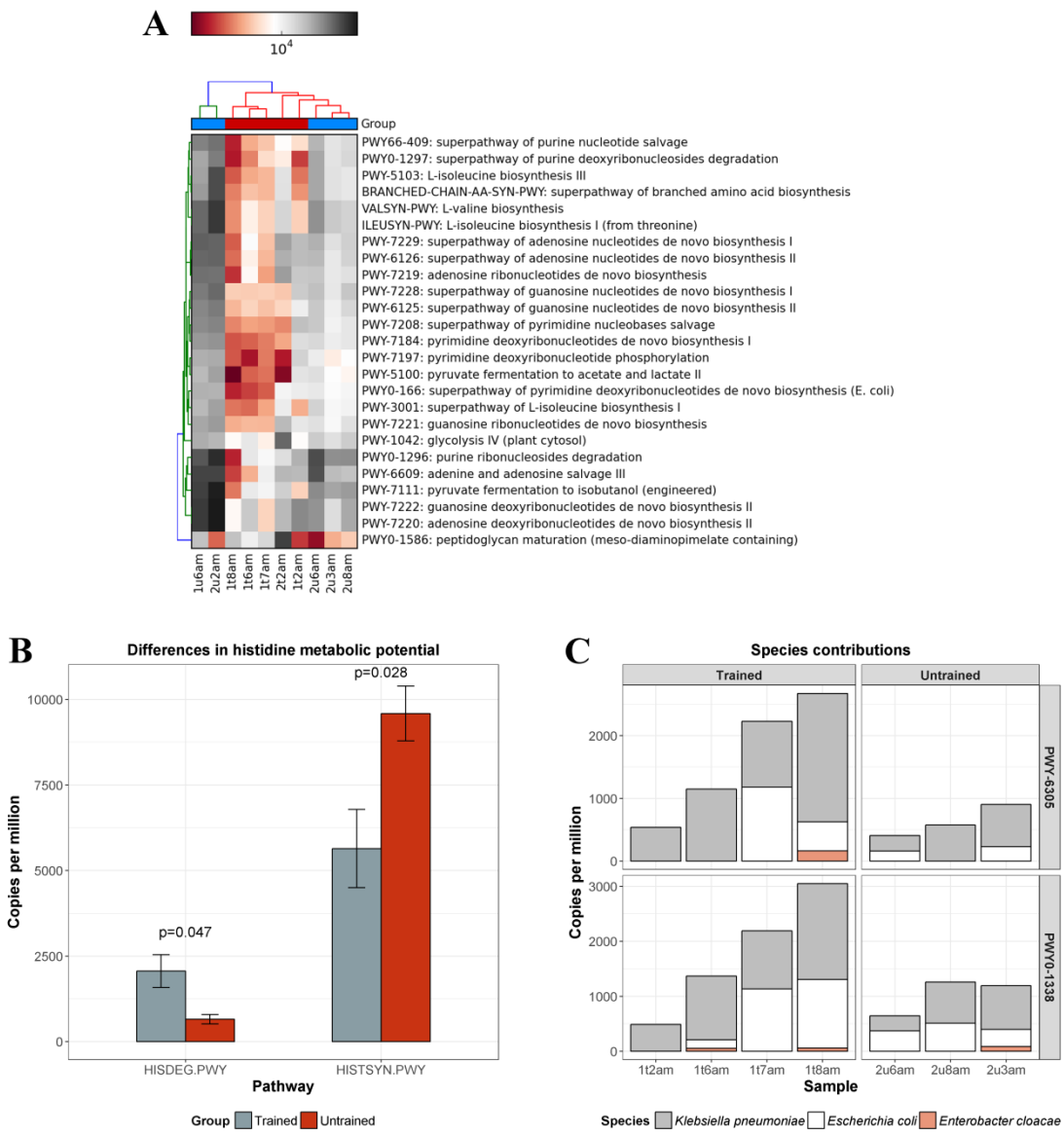


Figure 4: HUMAnN2 analysis. (A) Heat map showing the 25 most abundant MetaCyc pathways detected across the ten nunu metagenomic samples. (B) Bar plot showing differences in histidine metabolic potential between nunu samples from trained producers and nunu samples from untrained producers. (C) Bar plots showing the relative contributions of *E. cloacae, E. coli* and *K. pneumoniae* to the MetaCyc pathways PWY-6305 (putrescine biosynthesis) and PWY0-1338 (polymyxin resistance).

associated genes, including plasmid maintenance genes and transposition genes, in both groups.

Application of strain-level analysis to characterise enteric bacteria in nunu

Leonard *et al.* previously used metagenomic sequencing to detect *E. coli* in spinach which was intentionally spiked with *E. coli* O157:H7 strain Sakai (11). We downloaded the metagenomic reads from that study (16 samples) and we subjected them to StrainPhlAn, MetaMLST and PanPhlAn analysis, to confirm that these tools can accurately detect pathogens in food samples: MetaMLST was used for multi-locus sequence typing, StrainPhlAn was used for phylogenetic identification, and PanPhlAn was used for functional characterisation. MetaMLST accurately detected *E. coli* ST11 in 7/16 spinach samples (Table 1). StrainPhlAn detected *E. coli* strains in 5/16 samples and it showed that the *E. coli* strain in each of these samples was closely related to *E. coli* O157:H7 strain Sakai (Figure 5). PanPhlan detected Shiga toxin genes in 15/16 samples (Table 1) and it indicated that the *E. coli* strain in each of these samples was most closely related to *E. coli* O157:H7 strain Sakai. Thus, overall, PanPhlAn was the most sensitive method in this instance, since it was able to detect STEC in almost all of the samples, whereas the other tools detected STEC in less than half of the samples. In a follow-on study, Leonard *et al.* spiked spinach with 12 different Shiga toxin producing *E. coli* strains, and they detected single strains in 17 samples (18). We downloaded the metagenomic reads from the 17 samples and ran PanPhlAn, and were able to identify Shiga toxin genes in all 17 samples (Table S2).

Table 1: The results of MetaMLST and PanPhlAn analysis of spinach metagenomes spiked with *E. coli* O157:H7 Sakai

Sequence accession number	Reads	*E. coli* abundance (%)	stx2A	stx2B	Sequence type (ST)	Confidence (%)
SRR2177250	9,365,812	5.28412	1	1	Unknown	NA
SRR2177251	17,562,542	4.31712	1	1	11	99.97
SRR2177280	11,707,292	21.16364	1	1	100001	99.97
SRR2177281	10,580,532	2.84187	1	1	Unknown	NA
SRR2177282	6,155,636	60.51406	1	1	11	100
SRR2177283	13,120,244	10.11327	1	1	11	100
SRR2177284	7,500,056	2.05064	NA	NA	Unknown	NA
SRR2177285	14,482,370	66.69813	1	1	11	100
SRR2177286	14,035,970	69.17834	1	1	11	100
SRR2177287	12,242,348	5.62746	1	1	Unknown	NA
SRR2177288	8,303,788	10.75005	1	1	11	100
SRR2177357	14,621,672	8.02047	1	1	11	100
SRR2177358	10,684,052	3.18652	1	1	Unknown	NA
SRR2177359	4,964,436	1.17146	1	1	Unknown	NA
SRR2177360	12,729,834	1.81229	1	0	Unknown	NA
SRR2177361	11,946,092	0.70921	0	1	Unknown	NA

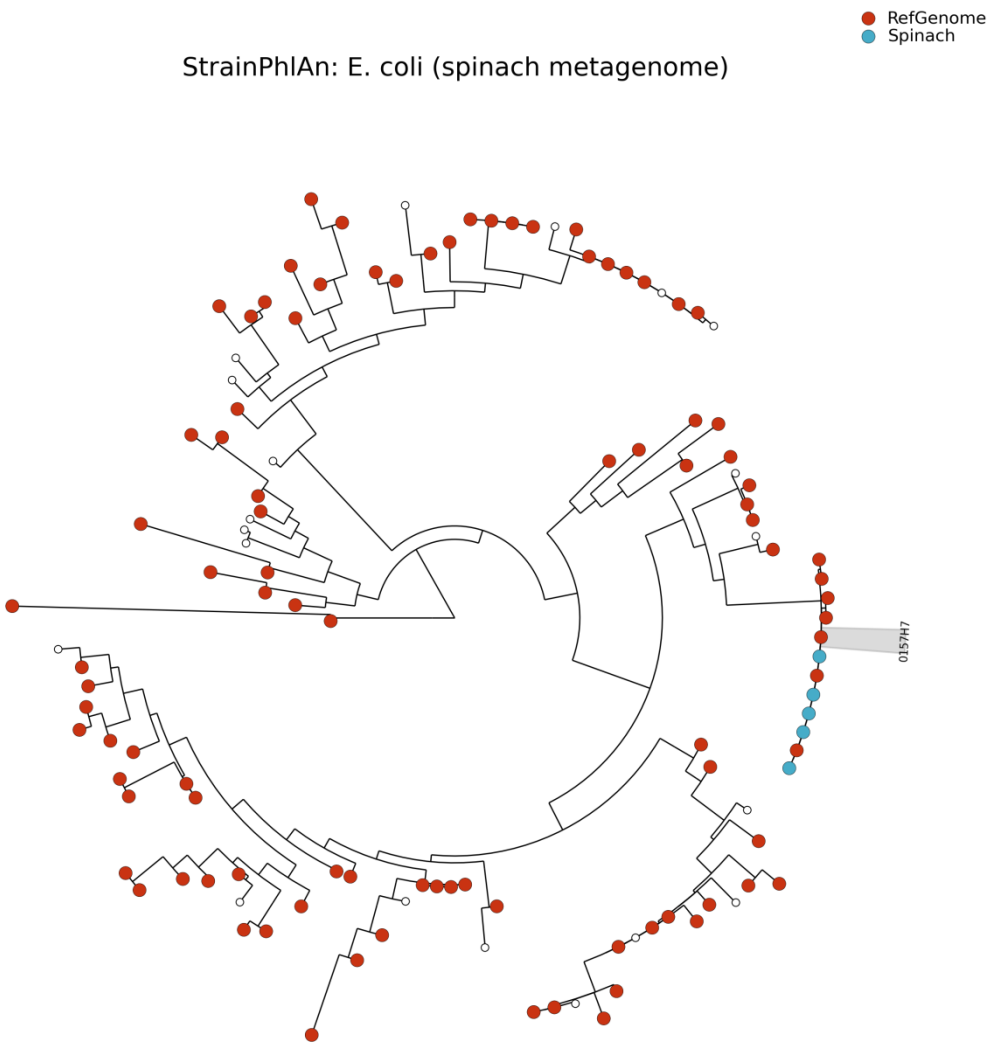


Figure 2: StrainPhlAn analysis of the spinach metagenome.

Having established the relative merits of these tools, we subsequently employed all three strategies to identify the strains of *E. coli* and *K. pneumoniae* present in the nunu samples. With regard to *E. coli,* MetaMLST detected a novel *E. coli* sequence type in 1t7am (Table 2). StrainPhlAn detected 24 *E. coli* marker genes in the samples and a phylogenetic tree (Figure 6a), which was generated by aligning these markers against 118 *E. coli* reference genomes (listed in Table S3), revealed that the *E. coli* strain in one sample, 1t7am, was closely related to *E. coli* O139:H28 E24377A. PanPhlAn detected *E. coli* strains in two samples: 1t7am and 1t8am. MDS analysis indicated that the strains from the two samples were functionally distinct from one another. Notably, a ShET2 enterotoxin encoding gene was identified in the *E. coli* strain from 1t7am. The same gene was found in *E. coli* O139:H28 E24377A. With regard to *K. pneumoniae,* MetaMLST detected the known sequence type *K. pneumoniae* ST39 in the sample 2u3am. Apparently novel *K. pneumoniae* sequence types were identified in six other samples (Table 1). StrainPhlAn detected 38 *K. pneumoniae* marker genes in the samples and a phylogenetic tree (Figure 6b), which was constructed by aligning these markers against 40 *K. pneumoniae* reference genomes (listed in Table S4), revealed that the *K. pneumoniae* strains in two samples, 1t8am and 2u3am, were closely related to *K. pneumoniae* KpQ3. In contrast, the *K. pneumoniae* strain in1t7am was most closely related to *K. pneumoniae* UCICRE 7. MDS analysis of the PanPhlAn output showed that five of the detected *K. pneumoniae* strains were functionally similar to one another (Figure 6c). However, two of the detected *K. pneumoniae* strains, in samples 1t6am and 1t7am, appeared to be functionally distinct from the others. In addition, PanPhlan indicated that sample 1t6am might have contained multiple strains, since an unusually high number of 5746 *K. pneumonia* gene families were detected. A TEM

Table 2: The results of MetaMLST analysis of the nunu metagenomic samples

Species	Sequence type (ST)	Confidence (%)	Sample
Klebsiella pneumoniae	100001	98.7	1t2am
Klebsiella pneumoniae	100002	100	1t6am
Esherichia coli	100001	100	1t7am
Klebsiella pneumoniae	100003	99.9	1t7am
Klebsiella pneumoniae	100004	100	1t8am
Klebsiella pneumoniae	39	100	2u3am
Klebsiella pneumoniae	100005	99.91	2u6am
Klebsiella pneumoniae	100006	99.91	2u8am

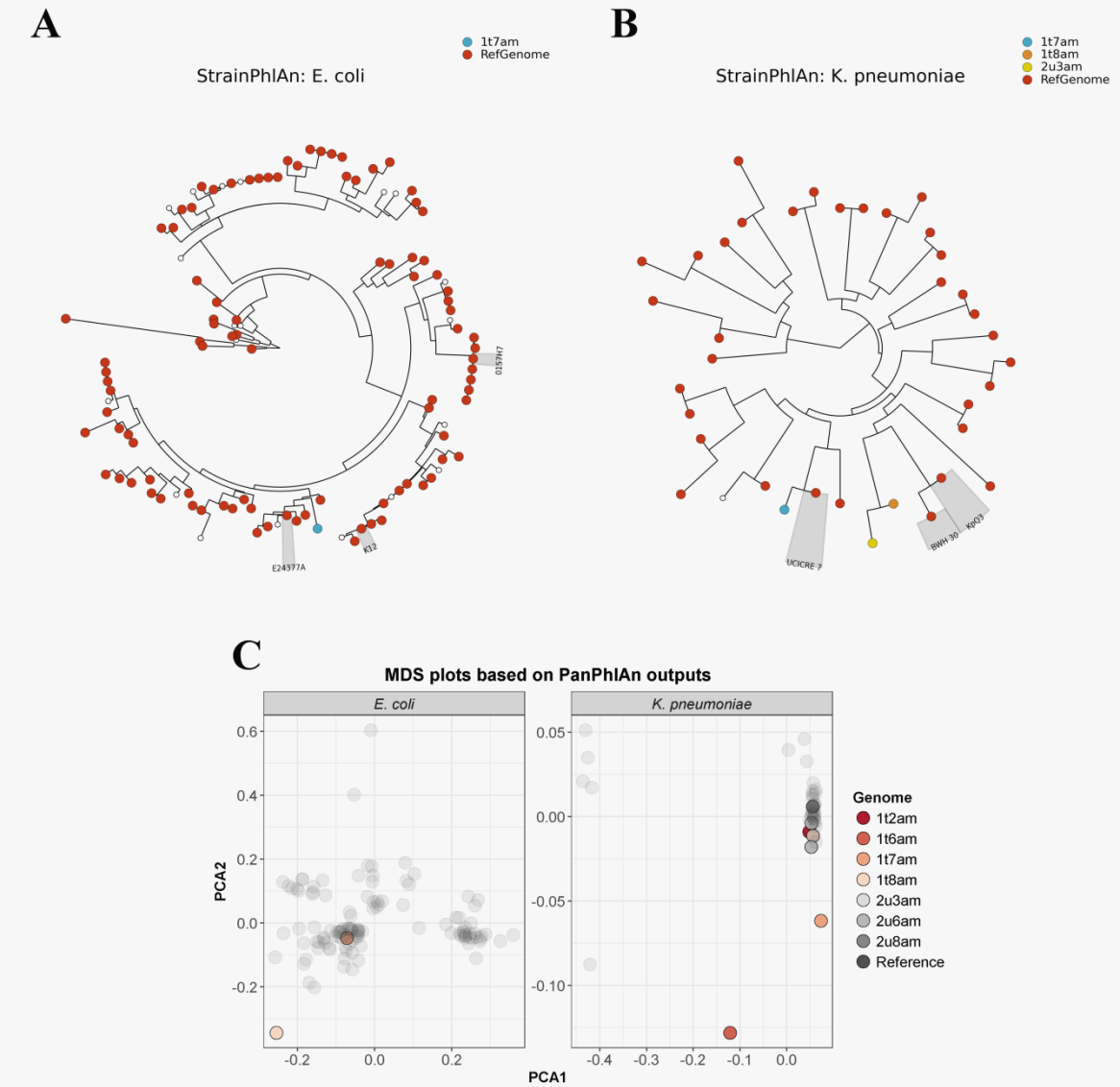


Figure 6: Strain-level analysis. Phylogenetic trees showing the relationships between (A) *E. coli* strains and (B) *K. pneumoniae* strains detected in the nunu metagenomic samples and their respective reference genomes, as predicted by StrainPhlAn. (C) MDS showing the functional similarities between strains detected in the nunu metagenomic samples, as predicted by PanPhlAn; reference genomes are shown in faded grey.

beta-lactamase gene was found in 1t2am using PanPhlAn and, furthermore, an OXA-48 carbapenemase gene was detected in 2u8am and the same gene was found in *K. pneumoniae* KpQ3.

Finally, we compared the time taken to process 10 nunu metagenome samples using the short-read alignment tools versus the metagenome assembler IDBA-UD (Figure S5). In each case, we observed that all of the short-read alignment tools were faster than IDBA-UD. It is important to note that additional bioinformatics analyses (contig binning, SNP analysis, etc.) are required to achieve strain-level identification from assembled metagenomes, and this emphasises the superior speed of the short-read alignment tools.

Discussion

Foodborne pathogens are responsible for millions of cases of disease annually, in the United States alone (43). High-throughput sequencing can potentially be used to detect pathogenic strains in food products to prevent the occurrence of disease outbreaks. A recent proof of concept study demonstrated that whole metagenome shotgun sequencing accurately detected Shiga toxin producing *E. coli* (STEC) strains in spiked spinach samples (18). However, that study used whole metagenome assembly-based approaches to achieve strain-level taxonomic resolution of the STEC in the samples. Whole metagenome assembly is a computationally intensive, time-consuming process, as illustrated by Nurk *et al.*, who recently reported that metagenome assembly can take between 1.5 hours to 6 hours, with a memory footprint ranging from 7.3 GB to 234.5 GB, to process a single human gut metagenomic sample, depending on the chosen assembler (44). Thus, the application of more rapid, less intensive bioinformatic tools for strain detection is desirable. In

this study, we demonstrate that the short read alignment-based programs MetaMLST, StrainPhlAn, and PanPhlAn can accurately identify pathogens in food products.

We validated the accuracy of each approach by processing spinach metagenome data from samples that were spiked with the STEC O157:H7 Sakai in a previous study (11). We observed that PanPhlAn was the most sensitive approach. Indeed, PanPhlAn was able to identify STEC in every sample where it was present at >2% relative abundance, whereas the other approaches worked best when STEC was present at high relative abundances. However, none of the tools detected *E. coli* O157:H7 Sakai in every sample tested. The observation of false negatives highlights that the tools are not entirely accurate. It is likely that increased sequencing depth and/or longer sequencing read lengths would reduce the false negative rate. We recommend that these tools be used to supplement data from metagenome sequence classifiers like MetaPhlAn2, which did detect *E. coli* in each sample. Therefore, we subsequently used the strain-level analysis tools in combination with other metagenomic approaches to assess the safety of nunu, a traditional Ghanaian fermented milk product.

Nunu is produced through the spontaneous fermentation of raw cow milk in calabashes or other containers for 24-36 hours at ambient temperature (23). The crude nature of the nunu production process has raised food safety concerns (25). Indeed, several potentially pathogenic microorganisms were previously detected in nunu samples by microbial culturing (25). This resulted in some nunu producers receiving hygiene practice training to improve food safety. However, our work suggests that there is little difference in the prevalence of pathogens in nunu samples from trained and untrained producers. One reason for this may be that it is difficult

for the nunu producers to adhere to the training recommendations which are not appropriate to the rural production conditions. During training, the producers were advised to pasteurise the milk before cooling and adding a starter culture. After incubating for 4-6 hours in a covered container, they were advised to stir the mixture and refrigerate the product. Lack of access to specific heat control and electricity, as well as the variance from the traditional method, which does not use a starter culture, are both reasons why the training is not adhered to.

16S rRNA gene sequencing revealed that the samples were dominated by Lactobacillales. However, we also detected high abundances of Enterobacteriales, including *Enterobacter* and *Escherichia,* in both groups. Subsequently, whole metagenome shotgun sequencing showed that most samples were dominated by *Streptococcus infantarius*, a species which was previously identified in other African dairy products (45, 46). Concernedly, *S. infantarius* has been linked to several human diseases, including bacteraemia (47), endocarditis (48) and colon cancer (49). Aside from *S. infantarius*, two other potentially pathogenic species, *Escherichia coli* and *Klebsiella pneumoniae,* were identified in a subset of samples.

Overall, our findings indicate that nunu samples from trained producers and untrained producers were contaminated with faecal material. Cattle faeces can be a major source of bacterial contaminants in raw cow milk (29), and thus, our results are not entirely surprising, but the remarkable abundance of such microorganisms in nunu is worrying. It had been hoped that nunu could be used to supplement traditional cereal-based weaning foods to improve infant nutrition. However, qualitative research among mothers and health workers highlighted safety concerns, which, as we have shown here, are valid. In particular, the presence of *E. coli* and *K.*

pneumoniae in nunu is a concern, and, thus, we employed strain-level metagenomics for the further characterisation of these bacteria.

In terms of *E. coli*, strain-level analysis indicated that the *E. coli* strain in one sample was an enterotoxin producer and it was closely related to *E. coli* O139:H28 E24377A, a strain which was linked to an outbreak of waterborne diarrhoea in India (50). In terms of *K. pneumoniae*, strain-level analysis indicated that the *K. pneumoniae* strains in two samples were antibiotic resistant and they were closely related to *K. pneumoniae* KpQ3, a strain which was linked to nosocomial outbreaks among burn unit patients. Thus, strain-level analysis suggests that there are likely pathogens in some of the samples. Interestingly, PanPhlAn also suggested that there were functionally distinct strains of both species in nunu samples from different producers. Perhaps, this indicates multiple incidences or sources of contamination. Undoubtedly, our work highlights an urgent need to further improve hygiene practices during nunu production, and the pasteurisation of the starting milk and the use of starter-based fermentation systems is an obvious solution.

In conclusion, our work suggests that short read alignment-based strain detection tools can be used to detect pathogens in other foods, apart from nunu or spinach, and they might also be useful for tracing the sources of foodborne disease outbreaks back to particular foods. Such tools are a significant improvement over 16S rRNA gene sequencing, which is often limited to genus-level identification, or metagenome read classification tools, which are limited to species-level identification (16). In addition, they are faster, and less computationally intensive, than metagenome assembly-based strain detection methods, making them more relevant to real-life scenarios which necessitate the rapid testing of many food samples. With DNA sequencing costs

continuing to decrease, the approach outlined here is an affordable option for food safety testing.

References

1. **Walsh AM, Crispie F, Claesson MJ, Cotter PD.** 2017. Translating Omics to Food Microbiology. Annual Review of Food Science and Technology **8**.
2. **Zheng J, Zhao X, Lin XB, Gänzle M.** 2015. Comparative genomics Lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. Scientific Reports **5:**18234.
3. **Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF.** 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nature Communications **6**.
4. **Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W.** 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104: H4 outbreak by rapid next generation sequencing technology. PLoS One **6:**e22751.
5. **Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A.** 2015. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing Escherichia coli O157. Clinical Infectious Diseases**:**civ318.
6. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP.** 2016. Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. Journal of Clinical Microbiology **54:**333-342.

7. **De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D.** 2016. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. Scientific Reports **6**.

8. **Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, Beresford TP, Ross RP, Fitzgerald GF, McSweeney PLH, Giblin L, Sheehan JJ, Cotter PD.** 2016. Thermus and the Pink Discoloration Defect in Cheese. mSystems **1**.

9. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.

10. **Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL.** 2016. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. Applied and Environmental Microbiology **82:**2433-2443.

11. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2015. Application of metagenomic sequencing to food safety: detection of Shiga toxin-producing Escherichia coli on fresh bagged spinach. Applied and Environmental Microbiology **81:**8183-8191.

12. **Wang Q, Garrity GM, Tiedje JM, Cole JR.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73:**5261-5267.

13. **Allard G, Ryan FJ, Jeffery IB, Claesson MJ.** 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinformatics **16:**1.

14. **Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML.** 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution **4:**1111-1119.

15. **Stellato G, Utter DR, Voorhis A, De Angelis M, Eren AM, Ercolini D.** 2017. A few Pseudomonas oligotypes dominate in the meat and dairy processing environment. Frontiers in Microbiology **8**.

16. **Lindgreen S, Adair KL, Gardner PP.** 2016. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports **6:**19233.

17. **Stasiewicz MJ, den Bakker HC, Wiedmann M.** 2015. Genomics tools in microbial food safety. Current Opinion in Food Science **4:**105-110.

18. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2016. Strain-Level Discrimination of Shiga Toxin-Producing Escherichia coli in Spinach Using Metagenomic Sequencing. PLoS One **11:**e0167870.

19. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

20. **Zolfo M, Tett A, Jousson O, Donati C, Segata N.** 2016. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Research**:**gkw837.

21. **Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N.** 2017. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems **2**.

22. **Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, Weir JC, Quince C, Smith GP, Betley JR.** 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104: H4. Jama **309:**1502-1510.

23. **Akabanda F, Owusu-Kwarteng J, Tano-Debrah K, Glover RL, Nielsen DS, Jespersen L.** 2013. Taxonomic and molecular characterization of lactic acid bacteria and yeasts in nunu, a Ghanaian fermented milk product. Food Microbiology **34:**277-283.

24. **Marsh AJ, Hill C, Ross RP, Cotter PD.** 2014. Fermented beverages with health-promoting potential: past and future perspectives. Trends in Food Science & Technology **38:**113-124.

25. **Akabanda F, Owusu-Kwarteng J, Glover R, Tano-Debrah K.** 2010. Microbiological characteristics of Ghanaian traditional fermented milk product, Nunu. Nature and Science **8:**178-187.

26. **Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, Claesson MJ.** 2016. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PLoS One **11:**e0148028.

27. **Schmieder R, Edwards R.** 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics **27:**863-864.

28. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.

29. **Doyle CJ, Gleeson D, O'Toole PW, Cotter PD.** 2017. Impacts of Seasonal Housing and Teat Preparation on Raw Milk Microbiota: a High-Throughput Sequencing Study. Applied and Environmental Microbiology **83:**e02694-02616.

30. **Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R.** 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics **26:**266-267.

31. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7:**335-336.

32. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of Molecular Biology **215:**403-410.

33. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research **41:**D590-D596.

34. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

35. **Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N.** 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods **12:**902-903.

36. **Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N.** 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ **3:**e1029.

37. **Silva GGZ, Green KT, Dutilh BE, Edwards RA.** 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics **32:**354-361.

38. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B.** 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Computational Biology **8:**e1002358.

39. **Peng Y, Leung HC, Yiu S-M, Chin FY.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28:**1420-1428.

40. **Team RC.** 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.

41. **McMurdie PJ, Holmes S.** 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One **8:**e61217.

42. **Fukushima M, Kakinuma K, Kawaguchi R.** 2002. Phylogenetic analysis of Salmonella, Shigella, and Escherichia coli strains on the basis of the gyrB gene sequence. Journal of Clinical Microbiology **40:**2779-2785.

43. **Scallan E, Hoekstra R, Mahon B, Jones T, Griffin P.** 2015. An assessment of the human health impact of seven leading foodborne pathogens in the

United States using disability adjusted life years. Epidemiology and Infection **143:**2795-2804.

44. **Nurk S, Meleshko D, Korobeynikov A, Pevzner PA.** 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research**:**gr. 213959.213116.

45. **Abdelgadir W, Nielsen DS, Hamad S, Jakobsen M.** 2008. A traditional Sudanese fermented camel's milk product, Gariss, as a habitat of Streptococcus infantarius subsp. infantarius. International Journal of Food Microbiology **127:**215-219.

46. **Jans C, Kaindi DWM, Böck D, Njage PMK, Kouamé-Sina SM, Bonfoh B, Lacroix C, Meile L.** 2013. Prevalence and comparison of Streptococcus infantarius subsp. infantarius and Streptococcus gallolyticus subsp. macedonicus in raw and fermented dairy products from East and West Africa. International Journal of Food Microbiology **167:**186-195.

47. **Beck M, Frodl R, Funke G.** 2008. Comprehensive study of strains previously designated Streptococcus bovis consecutively isolated from human blood cultures and emended description of Streptococcus gallolyticus and Streptococcus infantarius subsp. coli. Journal of Clinical Microbiology **46:**2966-2972.

48. **Herrero IA, Rouse MS, Piper KE, Alyaseen SA, Steckelberg JM, Patel R.** 2002. Reevaluation of Streptococcus bovis endocarditis cases from 1975 to 1985 by 16S ribosomal DNA sequence analysis. Journal of Clinical Microbiology **40:**3848-3850.

49. **Biarc J, Nguyen IS, Pini A, Gossé F, Richert S, Thiersé D, Van Dorsselaer A, Leize-Wagner E, Raul F, Klein J-P.** 2004. Carcinogenic

properties of proteins with pro-inflammatory activity from Streptococcus infantarius (formerly S. bovis). Carcinogenesis **25:**1477-1484.

50. **Tamhankar AJ, Nerkar SS, Khadake PP, Akolkar DB, Apurwa SR, Deshpande U, Khedkar SU, Stålsby-Lundborg C.** 2015. Draft genome sequence of enterotoxigenic Escherichia coli strain E24377A, obtained from a tribal drinking water source in India. Genome Announcements **3:**e00225-00215.

Supplemental material

Supplemental results

Bacterial culturing

Total counts were similar on plates incubated aerobically and anaerobically and but there was considerable variation between samples with counts ranging from lows of 10^7 bacteria/ml sample to highs of 10^{11}.

The selective culturing method indicated that more than 60% of the samples tested contained coliform bacteria, with a further 20% containing detectable *Staphylococcus*. The likelihood of culturing potentially pathogenic bacteria was the same in samples from trained and untrained producers.

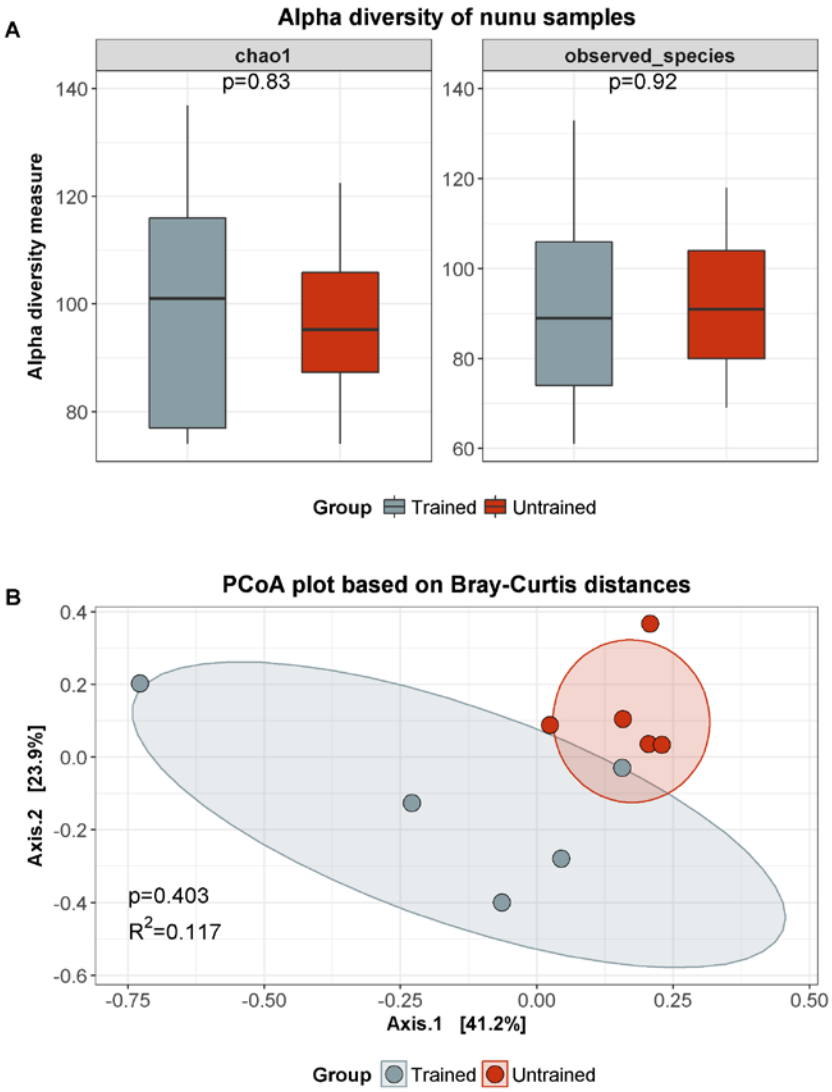


Figure S1: (A) Box plots showing the alpha diversity of nunu samples. (B) PCoA plot showing the beta diversity of nunu samples.

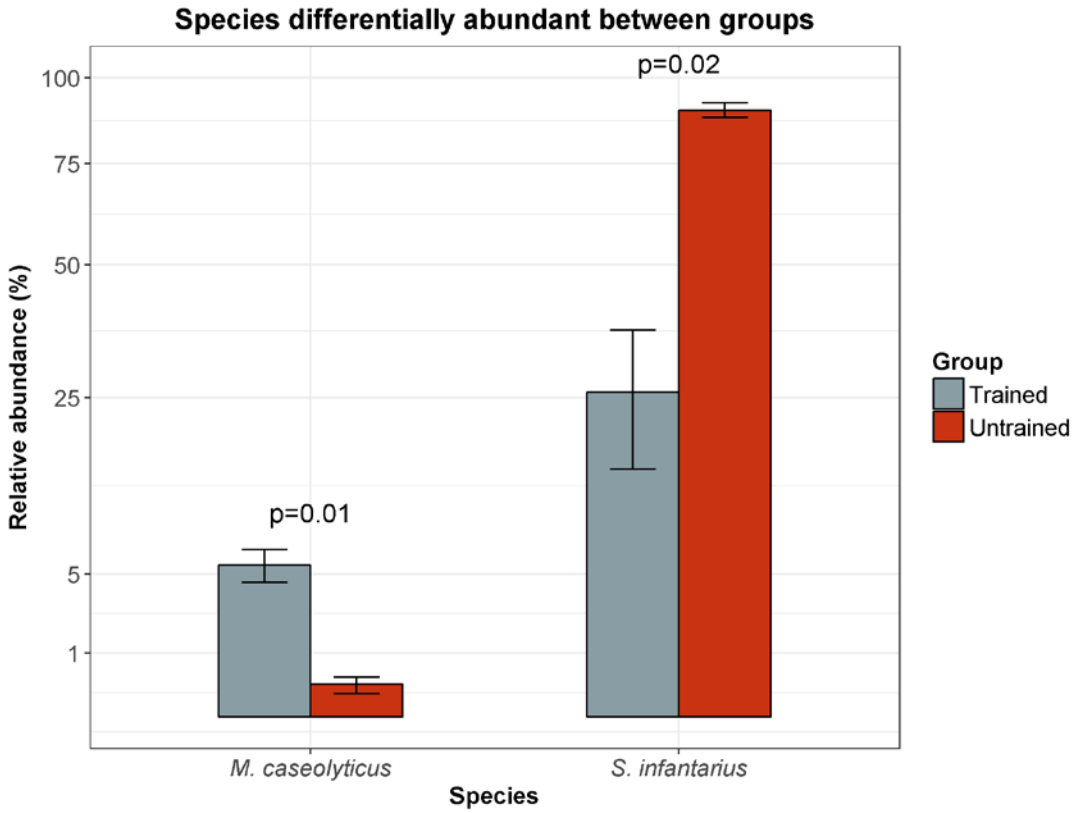


Figure S2: Bar plot showing species that were differentially abundant between nunu samples from trained producers and nunu samples from untrained producers.

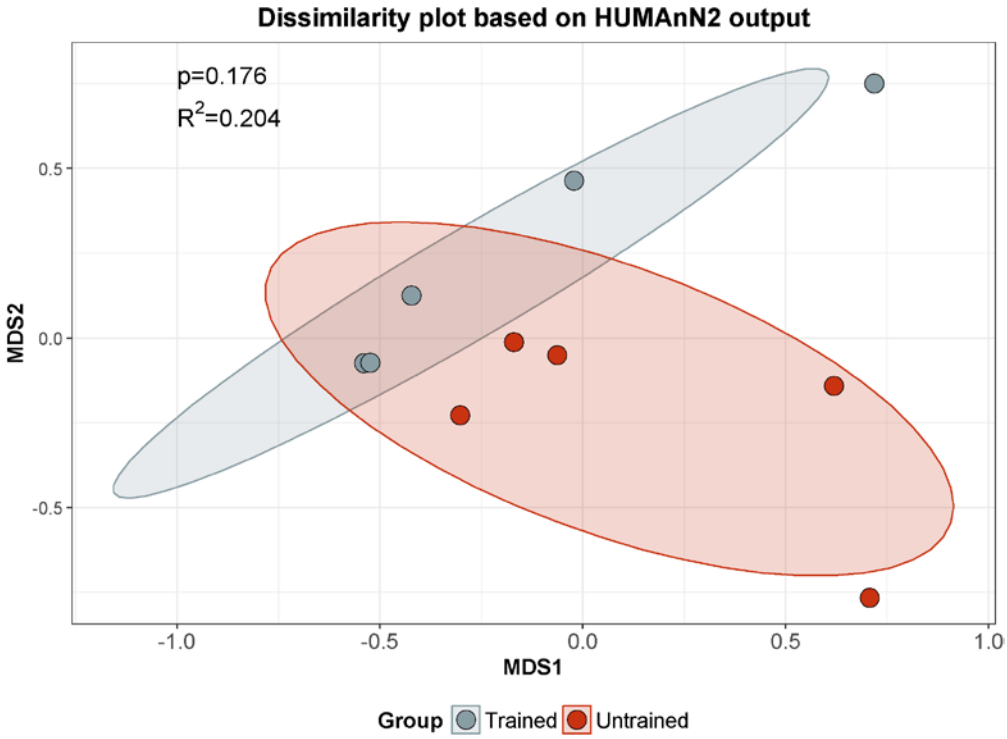


Figure S3: MDS plot showing the functional similarities between nunu samples from trained producers and nunu samples from untrained producers.

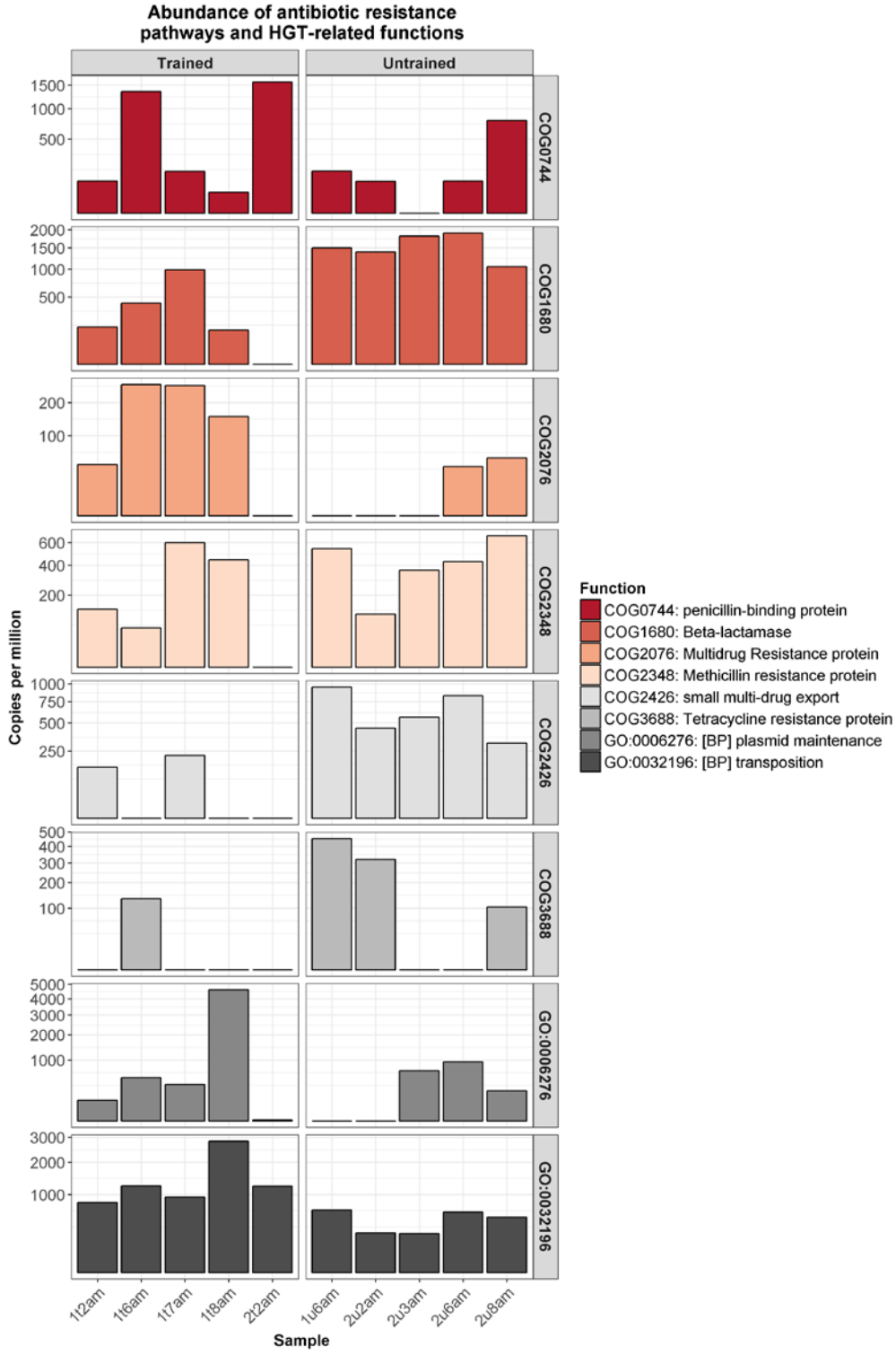


Figure S4: Bar plot showing the abundances of antibiotic resistance-associated functions and horizontal gene transfer (HGT)-associated functions in the nunu metagenome.

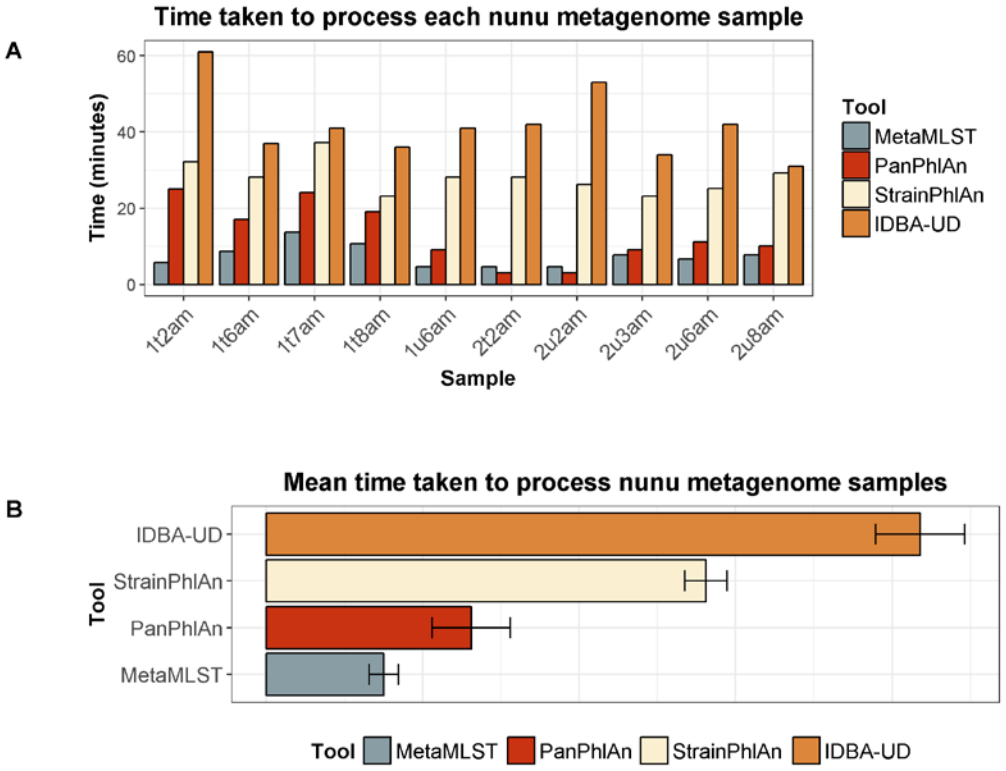


Figure S5: Bar plot showing (a) the total time taken to process nunu metagenomic samples, and (b) the mean time taken to process each nunu metagenomic sample, using IDBA-UD, MetaMLST, PanPhlAn and StrainPhlAn.

Table S1: MetaCyc pathways (detected by HUMAnN2) which were differentially different between trained versus untrained nunu samples

.

MetaCyc pathway	p-value (BH adjusted)
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis	0.016
ALLANTOINDEG-PWY: superpathway of allantoin degradation in yeast	0.019
ARGSYNBSUB-PWY: L-arginine biosynthesis II (acetyl cycle)	0.047
ARO-PWY: chorismate biosynthesis I	0.047
ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis	0.047
BRANCHED-CHAIN-AA-SYN-PWY: superpathway of branched amino acid biosynthesis	0.016
CALVIN-PWY: Calvin-Benson-Bassham cycle	0.009
COA-PWY-1: coenzyme A biosynthesis II (mammalian)	0.047
COA-PWY: coenzyme A biosynthesis I	0.016
COMPLETE-ARO-PWY: superpathway of aromatic amino acid biosynthesis	0.047
DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I	0.047
GALACTUROCAT-PWY: D-galacturonate degradation I	0.009
GLUTORN-PWY: L-ornithine biosynthesis	0.047
GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from ADP-D-Glucose)	0.016
HEME-BIOSYNTHESIS-II: heme biosynthesis I (aerobic)	0.028
HISDEG-PWY: L-histidine degradation I	0.047
HISTSYN-PWY: L-histidine biosynthesis	0.028
HOMOSER-METSYN-PWY: L-methionine biosynthesis I	0.028
ILEUSYN-PWY: L-isoleucine biosynthesis I (from threonine)	0.009
KDO-NAGLIPASYN-PWY: superpathway of (Kdo)2-lipid A biosynthesis	0.034
LACTOSECAT-PWY: lactose and galactose degradation I	0.009
NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch)	0.047
P122-PWY: heterolactic fermentation	0.028
P161-PWY: acetylene degradation	0.009
PENTOSE-P-PWY: pentose phosphate pathway	0.009
PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)	0.047
POLYISOPRENSYN-PWY: polyisoprenoid biosynthesis (E. coli)	0.047
PWY-2942: L-lysine biosynthesis III	0.016
PWY-3001: superpathway of L-isoleucine biosynthesis I	0.028
PWY-4242: pantothenate and coenzyme A biosynthesis III	0.047
PWY-5097: L-lysine biosynthesis VI	0.016
PWY-5100: pyruvate fermentation to acetate and lactate II	0.028
PWY-5103: L-isoleucine biosynthesis III	0.016
PWY-5104: L-isoleucine biosynthesis IV	0.034
PWY-5173: superpathway of acetyl-CoA biosynthesis	0.009
PWY-5265: peptidoglycan biosynthesis II (staphylococci)	0.016
PWY-5384: sucrose degradation IV (sucrose phosphorylase)	0.016
PWY-5686: UMP biosynthesis	0.028
PWY-5747: 2-methylcitrate cycle II	0.016
PWY-5850: superpathway of menaquinol-6 biosynthesis I	0.047

PWY-5860: superpathway of demethylmenaquinol-6 biosynthesis I	0.047
PWY-5913: TCA cycle VI (obligate autotrophs)	0.009
PWY-5973: cis-vaccenate biosynthesis	0.009
PWY-6124: inosine-5'-phosphate biosynthesis II	0.009
PWY-6125: superpathway of guanosine nucleotides de novo biosynthesis II	0.016
PWY-6147: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I	0.047
PWY-6163: chorismate biosynthesis from 3-dehydroquinate	0.028
PWY-6168: flavin biosynthesis III (fungi)	0.047
PWY-621: sucrose degradation III (sucrose invertase)	0.009
PWY-6282: palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)	0.047
PWY-6385: peptidoglycan biosynthesis III (mycobacteria)	0.028
PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)	0.028
PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)	0.028
PWY-6507: 4-deoxy-L-threo-hex-4-enopyranuronate degradation	0.009
PWY-6527: stachyose degradation	0.009
PWY-6901: superpathway of glucose and xylose degradation	0.028
PWY-6936: seleno-amino acid biosynthesis	0.009
PWY-7111: pyruvate fermentation to isobutanol (engineered)	0.009
PWY-7115: C4 photosynthetic carbon assimilation cycle, NAD-ME type	0.047
PWY-7184: pyrimidine deoxyribonucleotides de novo biosynthesis I	0.028
PWY-7187: pyrimidine deoxyribonucleotides de novo biosynthesis II	0.047
PWY-7197: pyrimidine deoxyribonucleotide phosphorylation	0.047
PWY-7199: pyrimidine deoxyribonucleosides salvage	0.009
PWY-7200: superpathway of pyrimidine deoxyribonucleoside salvage	0.015
PWY-7208: superpathway of pyrimidine nucleobases salvage	0.047
PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I	0.016
PWY-7242: D-fructuronate degradation	0.047
PWY-7357: thiamin formation from pyrithiamine and oxythiamine (yeast)	0.009
PWY-7539: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis III (Chlamydia)	0.047
PWY-7663: gondoate biosynthesis (anaerobic)	0.009
PWY0-1061: superpathway of L-alanine biosynthesis	0.047
PWY0-1296: purine ribonucleosides degradation	0.009
PWY0-1297: superpathway of purine deoxyribonucleosides degradation	0.009
PWY0-1298: superpathway of pyrimidine deoxyribonucleosides degradation	0.047
PWY0-1319: CDP-diacylglycerol biosynthesis II	0.009
PWY0-42: 2-methylcitrate cycle I	0.016
PWY66-409: superpathway of purine nucleotide salvage	0.009
PWY66-422: D-galactose degradation V (Leloir pathway)	0.047
RHAMCAT-PWY: L-rhamnose degradation I	0.047
SER-GLYSYN-PWY: superpathway of L-serine and glycine biosynthesis I	0.047
THRESYN-PWY: superpathway of L-threonine biosynthesis	0.028
TRPSYN-PWY: L-tryptophan biosynthesis	0.009
VALSYN-PWY: L-valine biosynthesis	0.009

Table S2: The results of PanPhlAn analysis of 17 spinach samples spiked with different STEC.

Sequence accession number	Strain	*E. coli* abundance (%)	stx1A	stx1B	stx2A	stx2B
SRR4101289	*E. coli* O157:H7 str. Sakai	89.73	1	1	1	1
SRR4101293	*E. coli* O157:H7 str. TW14359	79.9	1	1	1	1
SRR4101297	*E. coli* O157:H7 str. TW14359	42.74	0	0	1	1
SRR4101299	*E. coli* O113:H21 str. CL-3	45.7	0	0	1	1
SRR4101303	*E. coli* O113:H21 str. CL-3	68.17	1	1	0	0
SRR4101307	*E. coli* serovar O145:H28	92.98	0	0	1	1
SRR4101308	*E. coli* serovar O121:H19	92.14	0	0	1	1
SRR4101310	*E. coli* EC1738	60.59	0	0	1	1
SRR4101311	*E. coli* EC1738	87.5	0	0	1	1
SRR4101312	*E. coli* O104:H4 str. 2011C-3493	80.43	0	0	1	1
SRR4101314	*E. coli* O104:H4 str. 2011C-3493	66.08	0	0	1	1
SRR4101315	*E. coli* serovar O104:H7	89.98	0	0	1	1
SRR4101317	*E. coli* serovar O145:H28	20.56	0	0	1	1
SRR4101318	*E. coli* STEC_B2F1	38.67	0	0	1	1
SRR4101319	*E. coli* STEC_B2F1	56.95	0	0	1	1
SRR4101321	*E. coli* O113:H21 str. CL-3	92.83	0	0	1	1
SRR4101323	*E. coli* O113:H21 str. CL-3	76.2	0	0	1	1

Table S3: *Escherichia coli* reference genomes used in this study.

***Escherichia coli* strain**	**RefSeq assembly accession**
RM13514	GCF_000520035
ST540	GCF_000599625
ST2747	GCF_000599685
RM12761	GCF_000662395
RM12581	GCF_000671295
BIDMC 59	GCF_000692395
BIDMC 74	GCF_000692575
CHS 77	GCF_000692735
SE11	GCF_000010385
SE15	GCF_000010485
UTI89	GCF_000013265
536	GCF_000013305
APEC O1	GCF_000014845
E24377A	GCF_000017745
ATCC 8739	GCF_000019385
SMS-3-5	GCF_000019645
DH1	GCF_000023365
BL21-Gold(DE3)pLysS AG	GCF_000023665
IAI1	GCF_000026265
S88	GCF_000026285
UMN026	GCF_000026325
042	GCF_000027125
KO11	GCF_000147855
ABU 83972	GCF_000148365
UM146	GCF_000148605
MS 45-1	GCF_000164295
TA280	GCF_000176655
MS 145-7	GCF_000179115
W	GCF_000184185
LT-68	GCF_000188815
E1167	GCF_000190795
1.2741	GCF_000194175
3003	GCF_000194665
TW07793	GCF_000194685
UMNK88	GCF_000212715
96.0497	GCF_000215185
9.0111	GCF_000215265
UMNF18	GCF_000220005
STEC_DG131-3	GCF_000225125
clone D i14	GCF_000233895
B093	GCF_000242015
DEC2D	GCF_000249215
P12b	GCF_000257275

KO11FL	GCF_000258025
W	GCF_000258145
P4	GCF_000259425
APEC O78	GCF_000332755
KTE193	GCF_000351025
KTE233	GCF_000351325
KTE56	GCF_000351525
KTE66	GCF_000351625
KTE67	GCF_000351645
KTE17	GCF_000352125
KTE42	GCF_000352185
KTE29	GCF_000352245
KTE79	GCF_000352445
KTE84	GCF_000352465
KTE115	GCF_000352525
KTE135	GCF_000352585
KTE141	GCF_000352645
KTE144	GCF_000352665
KTE146	GCF_000352685
KTE147	GCF_000352705
KTE154	GCF_000352725
KTE192	GCF_000352785
KTE184	GCF_000352885
KTE183	GCF_000352905
KTE196	GCF_000352925
KTE197	GCF_000352945
KTE218	GCF_000353105
2720900	GCF_000355175
KTE114	GCF_000407765
KTE19	GCF_000407825
KTE31	GCF_000407925
KTE98	GCF_000408545
KTE102	GCF_000408585
HVH 55 (4-2646161)	GCF_000456825
HVH 58 (4-2839709)	GCF_000456865
HVH 65 (4-2262045)	GCF_000456945
HVH 111 (4-7039018)	GCF_000457555
HVH 115 (4-4465989)	GCF_000457655
HVH 139 (4-3192644)	GCF_000458035
HVH 164 (4-5953081)	GCF_000458495
HVH 188 (4-2356988)	GCF_000458825
HVH 195 (3-7155360)	GCF_000458955
KOEGE 44 (106a)	GCF_000459715
UMEA 3052-1	GCF_000460035
UMEA 3087-1	GCF_000460095
UMEA 3124-1	GCF_000460255

UMEA 3144-1	GCF_000460315
UMEA 3150-1	GCF_000460335
UMEA 3152-1	GCF_000460375
UMEA 3200-1	GCF_000460735
UMEA 3212-1	GCF_000460835
UMEA 3271-1	GCF_000461115
UMEA 3718-1	GCF_000461675
UMEA 4076-1	GCF_000461855
BIDMC 19C	GCF_000474825
JJ1886	GCF_000493755
HVH 36 (4-5675286)	GCF_000494935
K-12 substr. MG1655	GCF_000005845
12009	GCF_000010745
2009EL-2050	GCF_000299255
2009EL-2071	GCF_000299475
2011C-3493	GCF_000299455
11128	GCF_000010765
E2348/69	GCF_000026545
E24377A	GCF_000017745
EC4115	GCF_000021125
EDL933	GCF_000732965
Sakai	GCF_000008865
TW14359	GCF_000022225
Xuzhou21	GCF_000262125
11368	GCF_000091005
CB9615	GCF_000025165
RM12579	GCF_000245515
CE10	GCF_000227625
NRG 857C	GCF_000183345
55989	GCF_000026245
ETEC H10407	GCF_000210475

Table S4: *Klebsiella pneumoniae* reference genomes used in this study.

***Klebsiella pneumoniae* strain**	**RefSeq assembly accession**
HS11286	GCF_000240185
NTUH-K2044	GCF_000009885
KCTC 2242	GCF_000220485
Kp13	GCF_000512165
KPNIH31	GCF_000785005
234-12	GCF_000981845
DHQP1002001	GCF_001704235
Kp_Goe_154414	GCF_001902335
ATCC 13884	GCF_000163455
LCT-KP214	GCF_000255975
WGLW1	GCF_000300655
WGLW2	GCF_000300675
KpQ3	GCF_000300835
WGLW5	GCF_000300955
909957	GCF_000485755
BIDMC 40	GCF_000492215
BIDMC 36	GCF_000492295
BIDMC 25	GCF_000492315
BIDMC 24	GCF_000492335
BIDMC 23	GCF_000492355
UCICRE 14	GCF_000492415
UCICRE 7	GCF_000492535
BWH 30	GCF_000492695
BWH 28	GCF_000492735
MGH 44	GCF_000492795
MGH 43	GCF_000567685
XDR	GCF_000785625
KP-7	GCF_000406385
ATCC 25955	GCF_000409715
CCBH13327	GCF_000805735
-	GCF_000821685
ATCC 11296	GCF_000826585
50531633	GCF_001462885
YMC2010/8/B2027	GCA_001901745
12-3578	GCF_000367165
1183_KPNE	GCF_001060495
570_KPNE	GCF_001063755
k414	GCF_900085035
k2254	GCF_900085435
W2-15-ERG3	GCF_900093395

Chapter 6

Species classifier choice is a key consideration when analysing low complexity food microbiome data

Published in *Microbiome*

(doi: https://doi.org/10.1186/s40168-018-0437-0)

Authors: Aaron M. Walsh, Fiona Crispie, Orla O'Sullivan, Laura Finnegan, Marcus J. Claesson, and Paul D. Cotter.

Contributions:

- **Candidate** performed library preparations and computational analysis
- **OOS** provided guidance for bioinformatic analysis
- **LF** assisted in the sequencing library preparations
- **FC**, **MJC**, and **PC** supervised the study

Abstract

Background: The use of shotgun metagenomics to analyse low complexity microbial communities in foods has the potential to be of considerable fundamental and applied value. However, there is currently no consensus with respect to choice of species classification tool, platform or sequencing depth. Here, we benchmarked the performances of three high-throughput short-read sequencing platforms, the Illumina MiSeq, NextSeq 500, and Ion Proton, for shotgun metagenomics of food microbiota. Briefly, we sequenced six kefir DNA samples and a mock community DNA sample, the latter constructed by evenly mixing genomic DNA from 13 food-related bacterial species. A variety of bioinformatics tools were used to analyse the data generated, and the effects of sequencing depth on these analyses was tested by randomly subsampling reads.

Results: Compositional analysis results were consistent between the platforms at divergent sequencing depths. However, we observed pronounced differences in the predictions from species classification tools. Indeed, PERMANOVA indicated that there was no significant differences between the compositional results generated by the different sequencers ($p=0.693$, $R^2=0.011$), but there was a significant difference between the results predicted by the species classifiers ($p=0.001$, $R^2=0.127$). The relative abundances predicted by the classifiers, apart from MetaPhlAn2, were apparently biased by reference genome sizes. Additionally, we observed varying false-positive rates among the classifiers. MetaPhlAn2 had the lowest false-positive rate, whereas SLIMM had the greatest false-positive rate. Strain-level analysis results were also similar across platforms. Each platform correctly identified the strains present in the mock community, but accuracy was improved slightly with greater sequencing depth. Notably, PanPhlAn detected the dominant strains in each

kefir sample above 500,000 reads per sample. Again, the outputs from functional profiling analysis using SUPER-FOCUS were generally accordant between the platforms at different sequencing depths. Finally, and expectedly, metagenome assembly completeness was significantly lower on the MiSeq than either the NextSeq ($p=0.03$) or the Proton ($p=0.011$), and it improved with increased sequencing depth.

Conclusions: Our results demonstrate a remarkable similarity in the results generated by the three sequencing platforms at different sequencing depths, and, in fact, the choice of bioinformatics methodology had a more evident impact on results than the choice of sequencer did.

Background

Next generation sequencing has revolutionised microbiological research by enabling high-throughput metagenomic analysis of mixed microbial communities from many different environments (1-3). Briefly, metagenomics involves the culture-independent analysis of genomic DNA isolated from an entire microbial community, whereas genomics involves the culture-dependent analysis of genomic DNA isolated from a single microbial isolate (4). Metagenomic sequencing is an umbrella term which encompasses two distinct culture-independent sequencing approaches: amplicon sequencing or shotgun metagenomics. To date, amplicon sequencing, primarily of the 16S rRNA gene, has been the most commonly utilised metagenomics approach (5). 16S rRNA gene sequencing is used to investigate the bacterial composition of samples (6), but it is typically limited to genus-level identification (7), although higher resolution is sometimes possible (8, 9). In contrast, shotgun metagenomics enables species-level (10), and potentially strain-level classification (11-14) of microorganisms. Importantly, shotgun metagenomics can also be applied to determine the genetic content of samples to assess the associated functional potential (15). Shotgun metagenomics has been relatively underutilised, primarily because it is more expensive than 16S rRNA gene sequencing as it necessitates considerably higher sequencing depths (16). Indeed, desired sequencing depth is a factor that frequently dictates the choice of sequencing platform for high-throughput sequencing investigations (17).

A variety of sequencing platforms is currently available from several manufacturers, which vary in sequencing chemistry, read length and/or throughput. Presently, Illumina sequencers are the most commonly used sequencing platforms for microbiological research applications, including shotgun metagenomics (18).

Illumina sequencing chemistry is based on sequencing-by-synthesis, wherein adaptor-ligated DNA fragments on the surface of a flow cell are amplified by bridge PCR to generate clusters which are then sequenced via cyclic rounds of single-base extension with a mixture of fluorescently labelled dNTPs whose incorporation is detected using a high-sensitivity camera (19). The Illumina range of sequencers includes, in order of throughput, the MiSeq, NextSeq, and HiSeq series. Generally, the NextSeq or the HiSeq are preferred to the MiSeq for shotgun metagenomics, although there are several examples of the MiSeq also being used for this approach (20-22).

The Ion Torrent PGM from Life Technologies is another frequently utilised sequencer in microbiology, particularly for whole genome sequencing analysis of microbial isolates (23), although it is also used for shotgun metagenomics (24). In contrast, the higher-throughput Ion Proton, also from Life Technologies, is comparatively overlooked for metagenomic sequencing, whereas it is widely used for exome sequencing analysis of higher organisms (25-27). Ion sequencing chemistry is based on semiconductor sequencing, wherein adaptor-ligated DNA fragments attached to the surface of beads are amplified using emulsion PCR (28). Subsequently, these beads are placed inside microwells on a semiconductor sequencing chip, where a sequencing-by-synthesis reaction occurs which is similar to the Illumina method, except that base incorporation is determined by the measurement of pH changes caused by the escape of hydrogen ions during DNA extension.

Numerous studies have previously compared the performances of the Illumina MiSeq versus the Ion Torrent PGM to determine the relative accuracy of the sequencers and, now, it has been well established that the error rate of the Illumina

platforms, less than 1%, is lower than that of their Ion counterparts, approximately 1.7% (29). Specifically, Ion reads contain a higher incidence of insertions/deletions (30), and they are susceptible to premature sequence truncation (31). Long homopolymer tracts are especially problematic for Ion sequencing (32).

Previous investigations have aimed to determine if the choice of sequencing platform significantly influences metagenomic analyses. Recently, Fouhy *et al.* compared the MiSeq with the PGM for 16S rRNA gene sequencing analysis and reported that compositional results differed depending on the platform used (33). However, when these platforms were compared with the HiSeq for shotgun metagenomic applications, it was apparent that compositional results were similar across platforms but varied depending on the species classification tools used (34). Although these studies focused on gut microbial populations, shotgun metagenomics also has enormous potential with respect to the analysis of low complexity microbial communities, such as those in foods. Indeed, shotgun metagenomics has already vastly improved our knowledge of the microbiology of a number of fermented foods (35), and has numerous potential applications relating to food quality and safety (36). Furthermore, it has been proposed that metagenomic analysis of fermented foods can yield insights into the nature of microbial interactions or microbial community formation in other, more complicated, environments (37). However, the absence of a consensus with respect to the optimal sequencing platform or bioinformatic tools for shotgun metagenomic analysis of simple microbial communities could delay the more widespread application of the approach.

Here, we describe the first comparison of the performances of the short read DNA sequencing platforms, the Illumina MiSeq, the Illumina NextSeq, and the Ion Proton, for shotgun metagenomic-analysis of low complexity food-associated microbial

communities. This analysis was combined with an investigation of the impact of sequencing depth and downstream bioinformatic analysis, with a view to informing researchers, and especially food microbiologists, when designing shotgun metagenomic experiments.

Methods

Sources of metagenomic DNA

Metegenomic DNA representative of a low complexity, food-based, microbial community was generated by mixing equimolar ratios of genomic DNA from 13 food-related bacteria (Table 1). Strains were selected on the basis of the availability of corresponding complete or near-complete genome sequences from RefSeq (38). Genomic DNA was sourced from ATCC, DSM, and LMG. Genomic DNA concentration was determined prior to pooling using the Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland). We also analysed metagenomic DNA from six kefir milk samples which were previously isolated by Walsh *et al.* (39). Briefly, the samples were produced using either the Ick grain (samples: i24hd4; i24hd5; i24hd6) or the UK3 grain (samples: u24hd4; u24hd5; u24hd6). Three separate kefir fermentations were done using each grain. Fermented kefir samples were collected after 24 hours fermentation.

DNA sequencing

Illumina libraries were prepared using the Nextera XT kit in accordance with the Nextera XT DNA Library Preparation Guide from Illumina. MiSeq libraries were sequenced on the Illumina MiSeq sequencing platform in the Teagasc sequencing

facility, using a 2 x 300 cycle v3 kit, following standard Illumina sequencing protocols. NextSeq libraries were sequenced on the Illumina NextSeq 500, with a NextSeq 500/550 High Output Reagent Kit v2 (300 cycles), in accordance with standard Illumina sequencing protocols. Proton libraries were prepared in accordance with the Ion Xpress Plus gDNA Fragment Library Preparation User Guide. Proton libraries were enriched using the ION Proton PI template OT2 200 Kit v3, and sequenced using the Ion PI Sequencing 200 Kit v3, in accordance with standard Ion protocols.

Bioinformatic analysis

Raw shotgun metagenomic fastq files were converted to bam files using SAMtools (40), and duplicate reads were subsequently removed using Picard Tools (https://github.com/broadinstitute/picard). Next, low quality reads were removed using the trimBWAstyle.usingBam.pl script (https://github.com/genome/genome/blob/master/lib/perl/Genome/Site/TGI/Hmp/HmpSraProcess/trimBWAstyle.usingBam.pl). Specifically, Illumina reads were filtered to 200 bp, and reads with a quality score less than Q30 were discarded. Ion Proton reads were filtered to 110 bp, and reads with a quality score less than Q20 were discarded. The resulting fastq files were then converted to fasta files using the fq2fa option from IDBA-UD (41). Reads were randomly subsampled using seqtk (https://github.com/lh3/seqtk).

Compositional analysis was performed using the following species-classifiers: CLARK (42), Kaiju (43), Kraken (44), MetaPhlAn2 (45), and SLIMM (46). Species detected below 0.1% relative abundance were categorised as "other" for each

classifier. Note that Bowtie 2 (47) was used to map reads against the slimmDB_5k database. Strain-level metagenomic analysis was performed using PanPhlAn (12), which aligns reads against a pangenome database to functionally characterise strains. See supplemental material for a detailed description of the settings used for each species-classifier and/or PanPhlAn. Functional analysis was performed with SUPER-FOCUS (48), using the aligner DIAMOND (49), and HUMAnN2 (50), using the --bypass-translated-search option. Briefly, SUPER-FOCUS measures the abundances of subsystems, or groups of proteins with shared functionality, by aligning sequencing reads against a reduced SEED database (51), whereas HUMAnN2 measures the abundances of UniRef clusters (52) by aligning sequences against the ChocoPhlAn database. HUMAnN2 gene families were mapped to level-4 enzyme commission (EC) categories using HUMAnN2 utility mapping files. Metagenome assembly was performed using IDBA-UD (41).

Sequence data have been deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB22610.

Statistical analysis

Statistical analysis was performed in R-3.2.2 (53). The vegan package (version 2.3.0) (54) was used for alpha diversity analysis, as well as Bray-Curtis based multidimensional scaling (MDS) analysis. The adonis function in vegan was used for PERMANOVA (permutational analysis of variance) analysis, and the betadisper function, also in vegan, was used to calculate the distance of points from the centroid. The Kruskal-Wallis test was used to identify significant differences, and the resultant p-values were adjusted using the Benjamini-Hochberg method. The

Hmisc package (version 3.16.0) (55) was used for correlation analysis. The ggplot2 package (version 2.2.1) (56) was used for data visualisation.

It is important to note that the mock community DNA sample was only sequenced once on each platform, and thus we were unable to assess technical variation across sequencing runs. However, previous studies have already demonstrated that such variation is small, accounting for 1.3% to 2.3% variation between KEGG functional profiles (57). Additionally, we chose 0.1% relative abundance as an arbitrary cut-off to compare species or pathways, whereas, in reality, potentially important taxa or functions may be present below this threshold.

Results

Compositional analysis is influenced more by the choice of species-classifier than platform used

The Illumina MiSeq, the Illumina NextSeq, and the Ion Proton platforms were used for shogun metagenomic sequencing of a mock community sample, containing an equimolar mixture of genomic DNA from 13 food-related bacteria (Table 1), as well as six kefir DNA samples. The MiSeq produced 1,869,744 ± 401,024 reads per sample. The NextSeq produced 13,415,363 ± 4,098,763 reads per sample. The Proton produced 19,328,498 ± 3,240,112 reads per sample. The species classifiers CLARK, Kaiju, Kraken, MetaPhlAn2, and SLIMM were used to determine the bacterial composition of the samples. Compositional analysis of the mock community sample were generally consistent across the three platforms (Figure 1A), although some minor differences were observed, particularly between the Illumina sequencers versus the Ion Proton. For example, based on the average results from each species-classifier, the MiSeq, the NextSeq and the Proton detected *Acetobacter pasteurianus* in the mock community sample at 9.8%, 9.3% and 7.8%, respectively,

Table 1. Bacterial strains whose genomic DNA was mixed in an equimolar ratio to construct the Mock Community DNA sample.

Species	Strain	RefSeq assembly accession	GC content (%)	Genome size (bp)
Acetobacter pasteurianus	LMG 1513	GCF_000010825.1	53.1	2,907,495
Bacteroides vulgatus	DSM 1447	GCF_000012825.1	42.2	5,163,189
Bifidobacterium adolescentis Reuter	DSM 20083	GCF_000010425.1	59.3	2,089,645
Corynebacterium casei	LMG 19264	GCF_000550785.1	55.7	3,113,488
Gluconacetobacter medellinensis	LMG 1693	GCF_000182745.2	66.3	3,136,818
Lactobacillus brevis	ATCC 376	GCF_000014465.1	45.6	2,291,220
Lactobacillus casei	ATCC 334	GCF_000014525.1	46.6	2,895,264
Lactobacillus delbrueckii	DSM 20081*	GCF_001437195.1	49.7	415,890
Lactobacillus fermentum	LMG 18251	GCF_000010145.1	51.8	2,098,685
Lactobacillus reuteri	DSM 20016	GCF_000016825.1	38.9	1,999,618
Leuconostoc mesenteroides	LMG 6909*	GCF_000160595.1	37.7	543,364
Propionibacterium freudenreichii	LMG 16412	GCF_000940845.1	67.3	2,649,166
Streptococcus thermophilus	LMG 18311	GCF_000011825.1	39.1	1,796,846

* = incomplete genome sequence

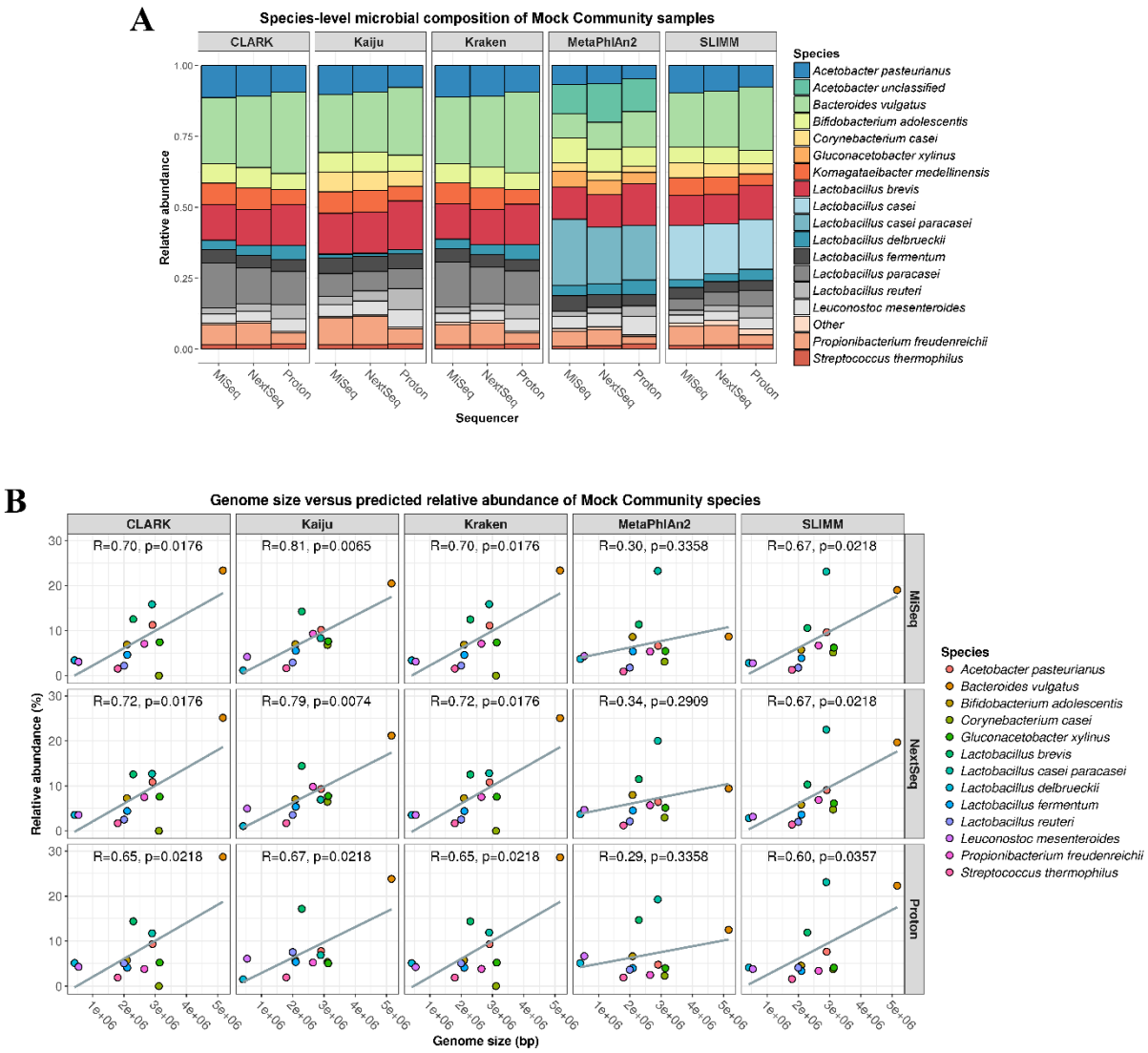


Figure 1: Compositional analysis of the mock community using the total number of reads from each sequencer. (A) Species-level profile of the mock community, as determined by each species-classifier. (B) Correlations between the relative abundances of species with their respective genome sizes.

and *Lactobacillus reuteri* in the same sample at 2.2%, 2.5% and 5.1%, respectively. With respect to species classifier, based on the average results from each sequencer, *Bacteroides vulgatus* was detected at 25.7% with CLARK compared to 10.2% with MetaPhlAn2, while *Lactobacillus brevis* was detected at 15.3% with Kaiju compared to 10.9% with SLIMM. Additionally, Kaiju, MetaPhlAn2, and SLIMM detected all 13 mock community species from data generated from each of the sequencing platforms used, whereas CLARK and Kraken did not detect *Corynebacterium casei* from any of the datasets, despite this species being represented with their respective databases. The mock community species were not present at equal relative abundances in any sample, despite genomic DNA having being mixed in equimolar ratios. For example, based on the average results from all data, the relative abundance of *Bacteroides vulgatus* was 20.8%, whereas the relative abundance of *Streptococcus thermophilus* was 1.6%. Indeed, the relative abundances of mock community species positively correlated with their genome size for all of the classifiers, apart from MetaPhlAn2 (Figure 1B). However, this observation is not entirely unexpected, since it is logical that larger reference genomes will receive more hits than smaller ones, and the issue has already been reported elsewhere (58). We subsequently found that normalising relative abundances, as predicted by CLARK, Kaiju, Kraken, and SLIMM, according to reference genome sizes resulted, on average, in a more equal distribution (Levene's test: $p=0.01$) (Figure S1). Note that since the *L. delbrueckii* DSM 20081 and *L. mesenteroides* LMG 6909 reference genomes were incomplete (Table 1), we normalised their abundances according to the median genome size for each species.

A number of species not present in the mock community DNA sample were detected as false-positives (Figure S2). With respect to platforms, the MiSeq and NextSeq

gave the lowest and highest numbers of false positives, respectively. Of the species classifiers, MetaPhlAn2 and SLIMM gave the lowest and highest numbers of false positives, respectively. However, it is important to note that all of the false positives were detected at less than 1% relative abundance, and species assigned were closely related to actual mock community species.

Overall, our results indicate that MetaPhlAn2 is the most accurate method, since it provided the lowest number of false positives. Additionally, the relative abundances predicted by MetaPhlAn2 were not biased by reference genome sizes.

The microbiota composition of kefir samples were similar as determined across the three platforms (Figure 2A), but again there were some significant differences. Specifically, two classifiers, Kaiju and SLIMM, indicated that *Lactobacillus plantarum* was present at significantly lower ratios in MiSeq-sequenced samples than Proton-sequenced samples (Kaiju: p=0.031; SLIMM: p=0.031), and SLIMM also indicated that *Lactobacillus acidophilus* was significantly lower in MiSeq samples compared to NextSeq samples (p=0.019). MetaPhlAn2 also failed to detect *Acetobacter* in MiSeq samples, but the tool did identify *Acetobacter* in the other sample groups. Alpha diversity measures were not significantly different between sequencers (Table S1), but they were significantly different between classifiers (Table S2). Specifically, the alpha diversity predicted by MetaPhlAn2 was lower than any other classifier, while the alpha diversity predicted by CLARK was also lower than SLIMM. Multidimensional scaling (MDS) analysis of compositional data confirmed that there was no significant dissimilarity between the sequencers (PERMANOVA: p= 0.693, R^2=0.011) (Figure 2B), but it revealed that there was a significant dissimilarity between the species classifiers (PERMANOVA: p=0.001, R^2=0.127) (Figure 2C). MetaPhlAn2 was especially different from the other

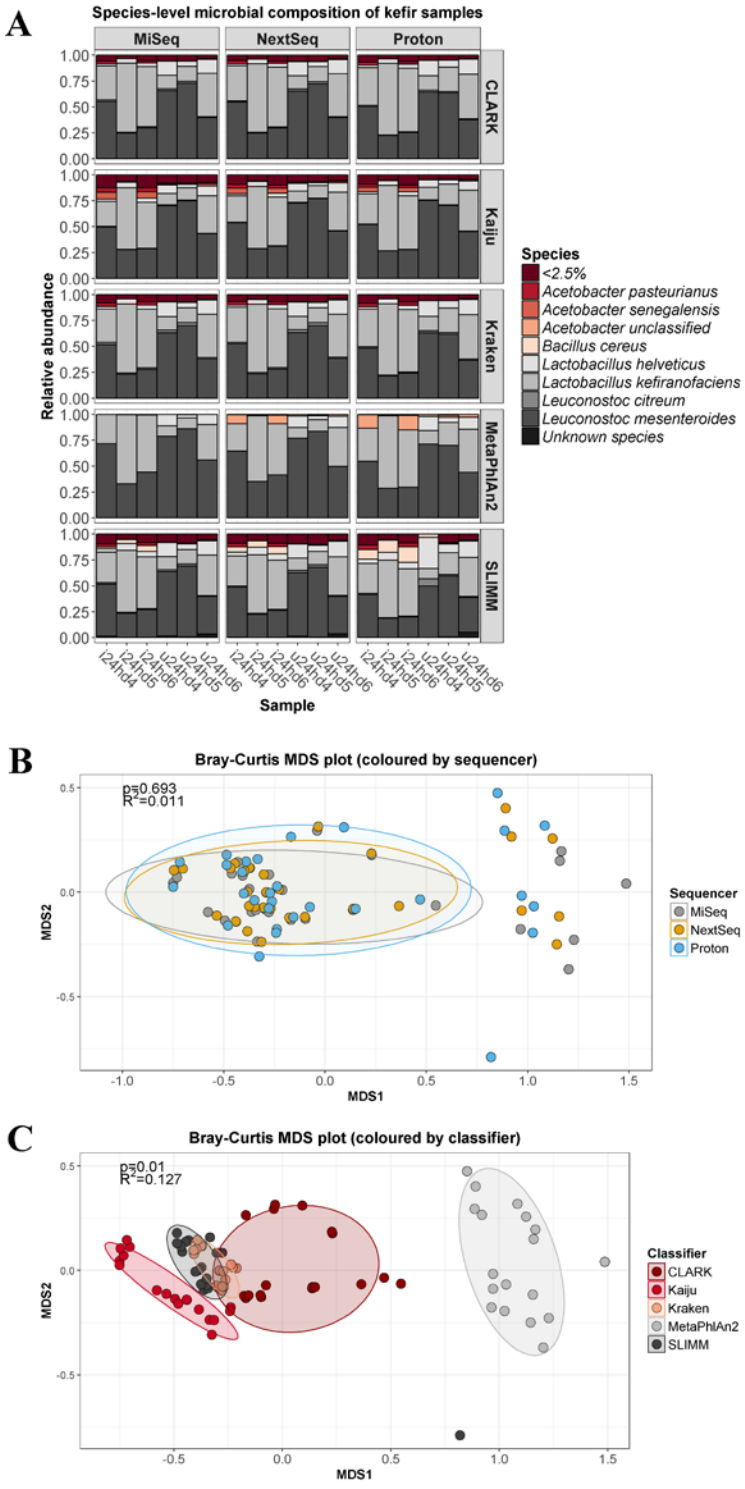


Figure 2: Compositional analysis of kefir samples using the total number of reads from each sequencer. (A) Species-level profile of the kefir samples, as determined by each species-classifier. (B) Dissimilarity plot showing differences between sequencers. (C) Dissimilarity plot showing differences between species-classifiers.

classifiers, since it did not detect *Acetobacter pasteurianus* or *Leuconostoc citreum* (Figure S3). Thus, although the mock community analysis indicated that MetaPhlAn2 is the most accurate approach, these results suggest that it is less sensitive than the other methods. Furthermore, only Kaiju detected *Acetobacter senegalensis*, while only SLIMM detected Bacillus cereus (Figure S3). However, there were no significant differences in the abundances of the two dominant kefir species, *Lactobacillus kefiranofaciens* or *Leuconostoc mesenteroides*, between any classifier (Table S3).

We averaged the results from each species classifier to generate a consensus taxonomic profile of the kefir samples (Figure S4A), and subsequent MDS analysis verified that there was no significant dissimilarity between the sequencers (PERMANOVA: $p=0.912$, $R^2=0.02$) (Figure S4B).

Bacterial strain identification was consistent across platforms

To further increase taxonomic resolution, we used PanPhlAn to characterise bacterial strains present in the samples. The results of strain-level metagenomic analyses were consistent across the three sequencers. For the mock community sample, PanPhlAn identified the correct strain of each of the analysed species (Figure 3A). For example, the MiSeq, NextSeq and Proton indicated that the *Lactobacillus fermentum* strain in the mock community shared 89.6%, 97.5% and 98.1%, respectively, of its pangenome gene-families with *L. fermentum* IFO 3956, while they indicated that the *Streptococcus thermophilus* strain in the mock community shared 76.6%, 86.9% and 96.7%, respectively, of its pangenome gene-families with *S. thermophilus* LMG 18311. Note that greater than two reference genomes are needed

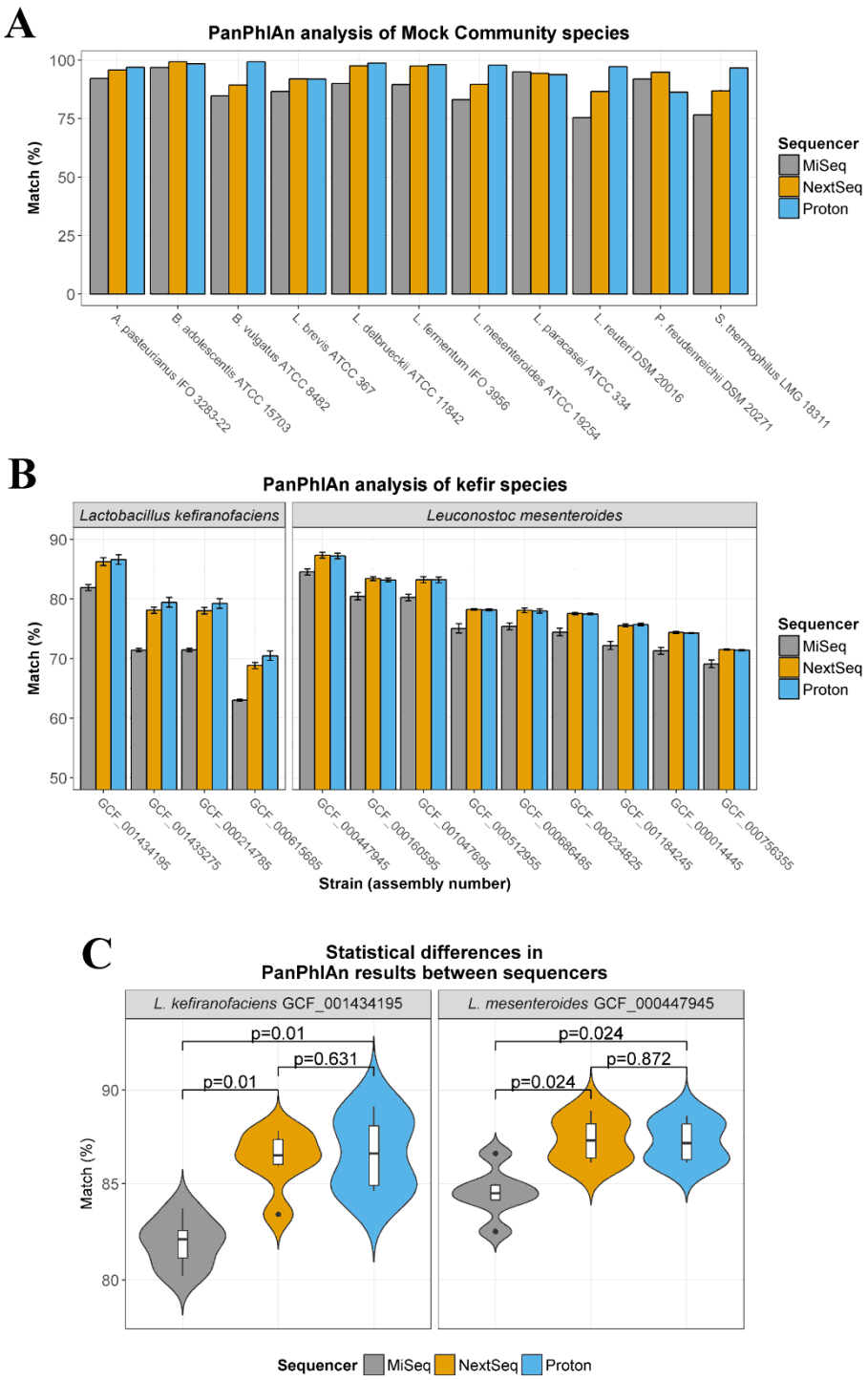


Figure 3: Strain-level analysis, with PanPhlAn, using the total number of reads from each sequencer. (A) The highest match for each of 11 mock community species for which ≥2 reference strain genomes are available at RefSeq, based on the presence/absence of pangenome gene-families. (B) A comparison of the relatedness of the *Lactobacillus kefiranofaciens* and *Leuconostoc mesenteroides* strains detected in kefir samples with each of the reference strain genomes present in the respective PanPhlAn pangenome databases.

to construct a PanPhlAn pangenome database, and hence we were unable to use PanPhlAn for strain-level analysis of *Corynebacterium casei* or *Gluconacetobacter xylinus.*

For the kefir samples, PanPhlAn was used to provide strain-level analysis of the two most dominant species, *Lactobacillus kefiranofaciens* and *Leuconostoc mesenteroides*. Analysis on the MiSeq, NextSeq and Proton platforms all indicated that the *Lactobacillus kefiranofaciens* strain detected in the kefir samples was most closely related to *L. kefiranofaciens* GCF_001434195, but the MiSeq detected significantly fewer shared pangenome gene-families than either the NextSeq ($p=0.01$) or the Proton ($p=0.01$). Similarly, analysis of data from all three platforms indicated that the *Leuconostoc mesenteroides* strain was most closely related to *L. mesenteroides* GCF_000447945 (Figure 3B), but, again, the MiSeq detected significantly fewer shared pangenome gene-families than either the NextSeq ($p=0.024$) or the Proton ($p=0.024$). It is likely that the decreased accuracy achieved with the MiSeq was due to its lower sequencing depth relative to the other two sequencers. The contribution of sequencing depth to the accuracy of strain-level analysis is investigated in subsequent sections.

Metagenome assembly completeness varies significantly between platforms but functional profiles remain consistent

IDBA-UD was used to assemble the mock community and kefir metagenomes. The n50 number, which is a measure of metagenome assembly completeness, of MiSeq assembles was significantly lower than either NextSeq ($p=0.03$) or Proton assemblies

(p=0.011) (Figure S5). The mean n50 numbers for each platform were as follows: n50=3,151 (MiSeq); n50=13,874 (NextSeq); and n50=9,307 (Proton).

The functional profile of the mock community sample, as characterised by SUPER-FOCUS, was congruent across the three platforms (Figure 4A). As anticipated, a large proportion of the metagenome was involved in housekeeping functions such as carbohydrate or protein metabolism. Specifically, the MiSeq, NextSeq and Proton detected the "carbohydrates" subsystem at 18.2%, 18.4% and 18.7%, respectively, while they detected the "protein metabolism" subsystem at 8.4%, 8.3% and 8.4%, respectively. Similarly, the functional potential of kefir samples was accordant across the three platforms. Indeed, MDS analysis indicated that the Illumina sequencers were more similar to each other than the Proton, but there was no significant overall dissimilarity between the three sequencers (PERMANOVA: p= 0.808, R^2=0.057) (Figure 4B). However, we did observe significant differences in the abundances of three SUPER-FOCUS subsystems that were present at greater than 1% relative abundances in kefir. Specifically, assignments to the "fatty acid" subsystem was significantly higher among the samples sequenced on the MiSeq than those sequenced with the NextSeq (p=0.049); levels of "heat shock" subsystem-assigned reads were significantly different between all three platforms (MiSeq versus NextSeq: p=0.01; MiSeq versus Proton: p=0.037; NextSeq versus Proton: p=0.01); and reads assigned to the "protein biosynthesis" subsystem were significantly higher among samples sequenced on the Proton than those sequenced with either the MiSeq (p=0.037) or the NextSeq (p=0.037) (Figure 4C).

Metagenomic pathway analysis tools provide inconsistent results

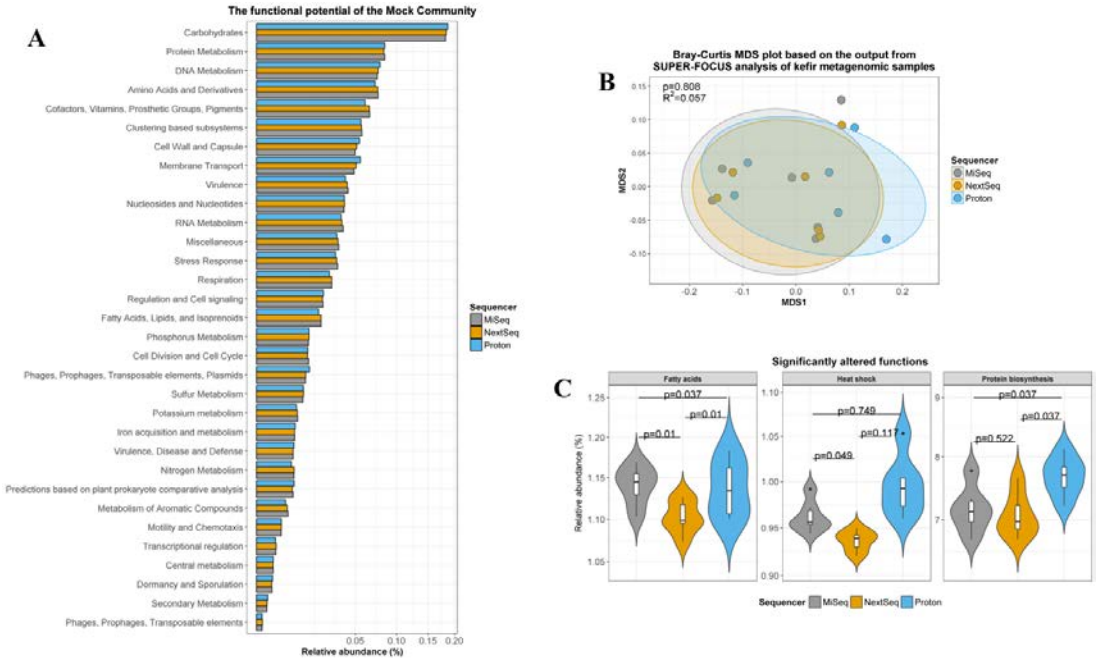


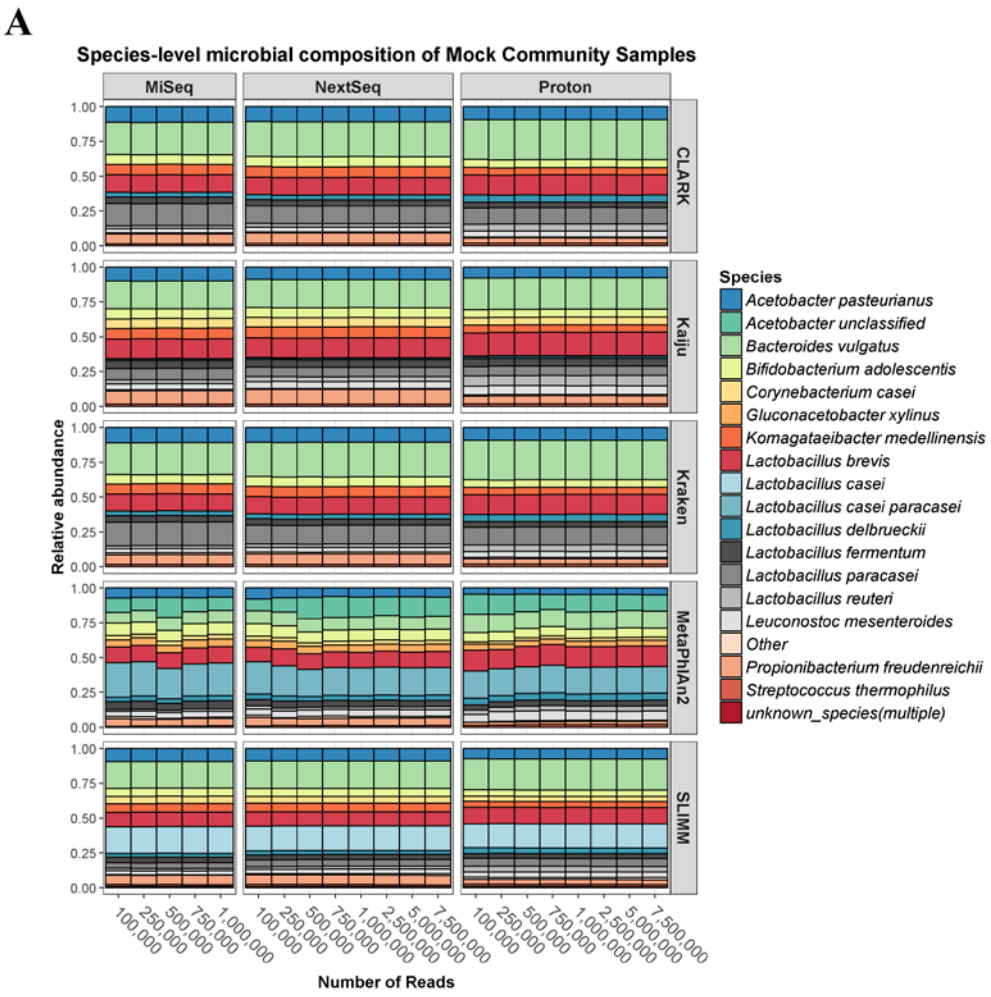
Figure 4: Functional analysis, with SUPER-FOCUS, using the total number of sequences from each sequencer. (A) The relative abundances of SUPER-FOCUS level-1 subsystems detected in the mock community. (B) Dissimilarity plot based on the relative abundances of the SUPER-FOCUS level 3 subsystems detected in the kefir samples. (C) SUPER-FOCUS level 2 subsystems which were significantly altered between sequencers.

The results from SUPER-FOCUS were compared to those from HUMAnN2, which is an alternative tool for functional analysis of metagenomes. MDS analysis revealed that there was a significant dissimilarity between the two tools (PERMANOVA: p= 0.808, R^2=0.057) (Figure S6), based on the relative abundances of 865 level-4 enzyme commission (EC) categories which were detected by both programs. Indeed, in total, 749 EC categories were differentially abundant between the methods.

Sequencing depth does not significantly affect measured composition or predicted functional potential of low complexity food microbiomes

Reads from the mock community and kefir samples were randomly subsampled to assess the effects of sequencing depth on compositional and functional analysis. MiSeq reads were subsampled from 100,000 to 1,000,000 reads per sample, while NextSeq and Proton reads were subsampled from 100,000 to 7,500,000 reads per sample.

For the mock community sample, the compositions were close to identical, regardless of sequencing depth (Figure 5A). For example, Kraken detected *Lactobacillus reuteri* at 2.6% using 100,000 NextSeq reads, while it was detected at 2.5% using 7,500,000 NextSeq reads. Similarly, the results of compositional analysis were uniform at divergent sequencing depths (Figure 5B). For instance, based on SUPER-FOCUS results, the carbohydrate metabolism subsystem was detected at 18.6% using 100,000 NextSeq reads, while it was detected at 18.4% using 7,500,000 NextSeq reads.



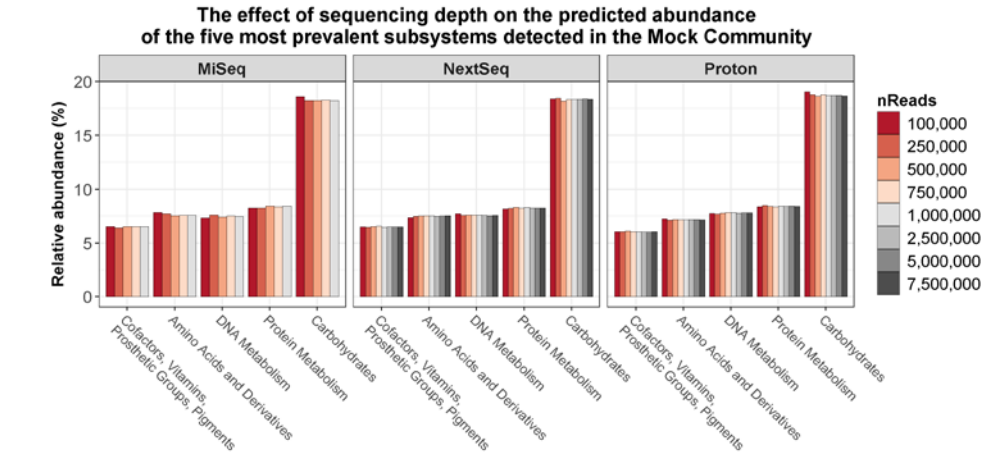


Figure 5: The effect of sequencing depth on compositional and functional analysis of the mock community. (A) The species-level profile of the mock community sample at different sequencing depths on each sequencer. (B) The relative abundances of the top five most prevalent SUPER-FOCUS level 1 subsystems detected in the mock community at different sequencing depths on each sequencer.

The microbial profiles of the subsampled kefir reads were highly similar at different sequencing depths (Figure 6A). Indeed, there were no significant differences in the abundances of any species present at >0.1% relative abundance, as detected by each classifier, at sequencing depths of 100,000, 1,000,000 or 7,500,000 reads per sample. However, we did observe some notable, albeit non-significant, differences (Figure 6B). Specifically, MetaPhlAn2 indicated that the abundance of *Acetobacter* was lower at 100,000 NextSeq reads compared to 7,500,000 NextSeq reads (p=0.06). SLIMM indicated that the abundance of *Latcobacillus casei* was lower at: 100,000 MiSeq reads compared to 1,000,000 MiSeq reads (p=0.054); 100,000 NextSeq reads compared to 7,500,000 NextSeq reads (p=0.056); and 1,000,000 NextSeq reads compared to 7,500,000 NextSeq reads (p=0.056). Additionally, there were no significant differences in alpha diversity at these different sequencing depths on any sequencer (Table S5), although alpha diversity measures predicted by MetaPhlAn2 did visibly increase with sequencing depths up to 1,000,000 reads per sample (Figure S7A). Similarly, MDS analysis indicated that there were no clear differences in microbial composition predicted by CLARK, Kaiju, Kraken or SLIMM at different sequencing depths, but there were apparent differences between the microbial compositions predicted by MetaPhlAn2 at different sequencing depths (Figure S7B). It is important to note that we only included species which were detected at >0.1% relative abundance in our diversity analysis. It is possible that higher sequencing depths might improve the detection of species present at <0.1%, which may affect diversity measures.

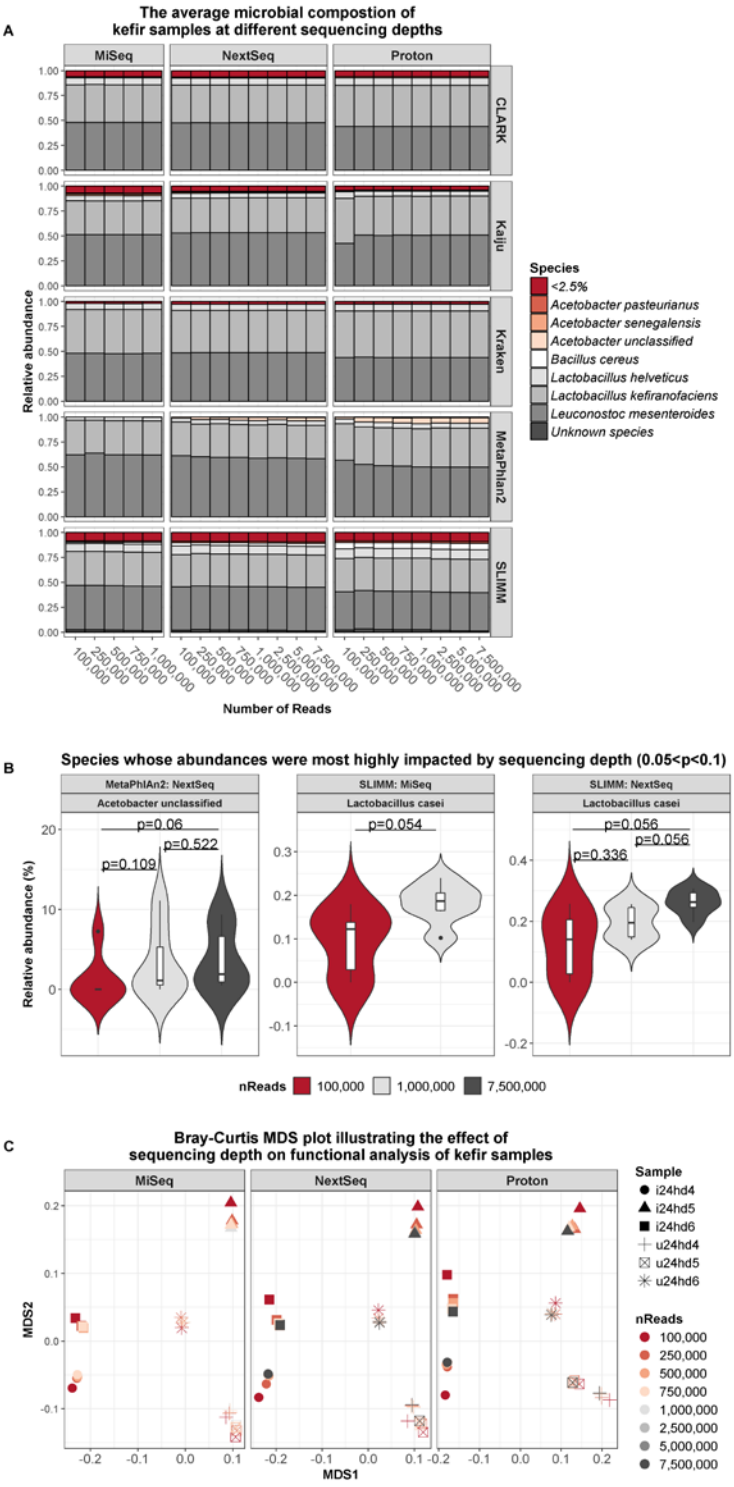


Figure 6: The effect of sequencing depth on compositional and functional analysis of kefir. (A) The average species-level profile of kefir samples at different sequencing depths on each sequencer. (B) Species whose abundances were most highly impacted by sequencing depth (0.05<p<0.1). (C) Dissimilarity plot based on the relative abundances of the SUPER-FOCUS level 3 subsystems detected in the kefir samples at different sequencing depths on each sequencer.

SUPER-FOCUS analysis of subsampled kefir reads again revealed that the functional profiles were highly similar at the different sequencing depths. Indeed, MDS analysis indicated that data points did not cluster by the number of reads per sample (Figure 6C), but instead we identified six distinct clusters, representing each of the six kefir samples. However, we did identify fifteen differentially abundant level 2 subsystems at different sequencing depths, but these functions were all present at <0.01% relative abundance (Figure S8).

Metagenome assembly of subsampled kefir reads using IDBA-UD showed that sequencing depth had a major impact on metagenome completeness (Figure 7A). The n50 number of metagenomes assembled from 100,000 reads was significantly lower than the n50 number of those assembled from 1,000,000 reads (p=0.003) or 7,500,000 reads (p=0.003) (Figure 7B). Additionally, the n50 number of metagenomes assembled from 1,000,000 reads was significantly lower than the n50 number of those assembled from 7,500,000 reads (p=0.009).

Finally, we used PanPhlAn to assess the impact of sequencing depth on strain-level analysis of the two dominant kefir species, *L. kefiranofaciens* and *L. mesenteroides*. Below 500,000 reads per sample, PanPhlAn failed to characterise either species at the strain-level for several kefir samples on each sequencer, but above 500,000 reads per sample, PanPhlAn successfully characterised both species at the strain-level for every kefir sample on each sequencer (Figure 8A). PanPhlAn indicated that the *L. kefiranofaciens* and *L. mesenteroides* strains detected in kefir samples shared the greatest similarity to *L. kefiranofaciens* GCF_001434195 and *L. mesenteroides* GCF_000447945, respectively. However, the percentage shared pangenome gene-families was significantly lower at 500,000 reads per sample compared to 7,500,000 reads per sample on the NextSeq for both species (*L.* kefiranofaciens: p=0.031; *L.*

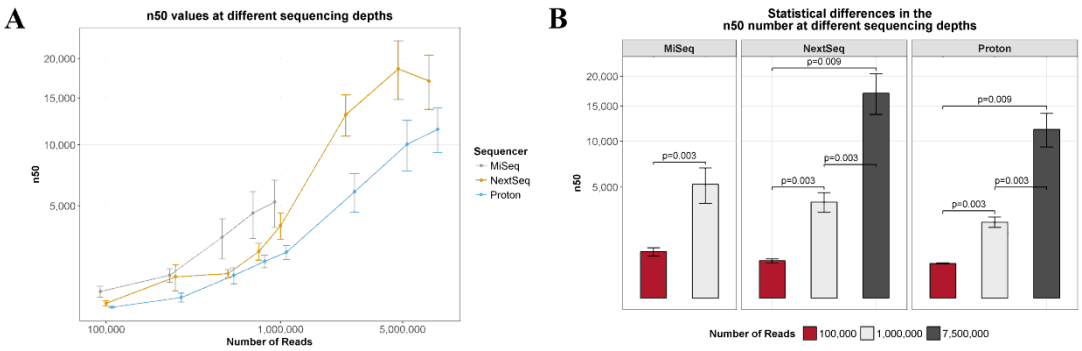


Figure 7: The effect of sequencing depth on metagenome assembly using IDBA-UD. (A) The n50 numbers at each sequencing depth. (B) Statistical differences in the n50 number at 100,000, 1,000,000 and 7,500,000 reads per sample.

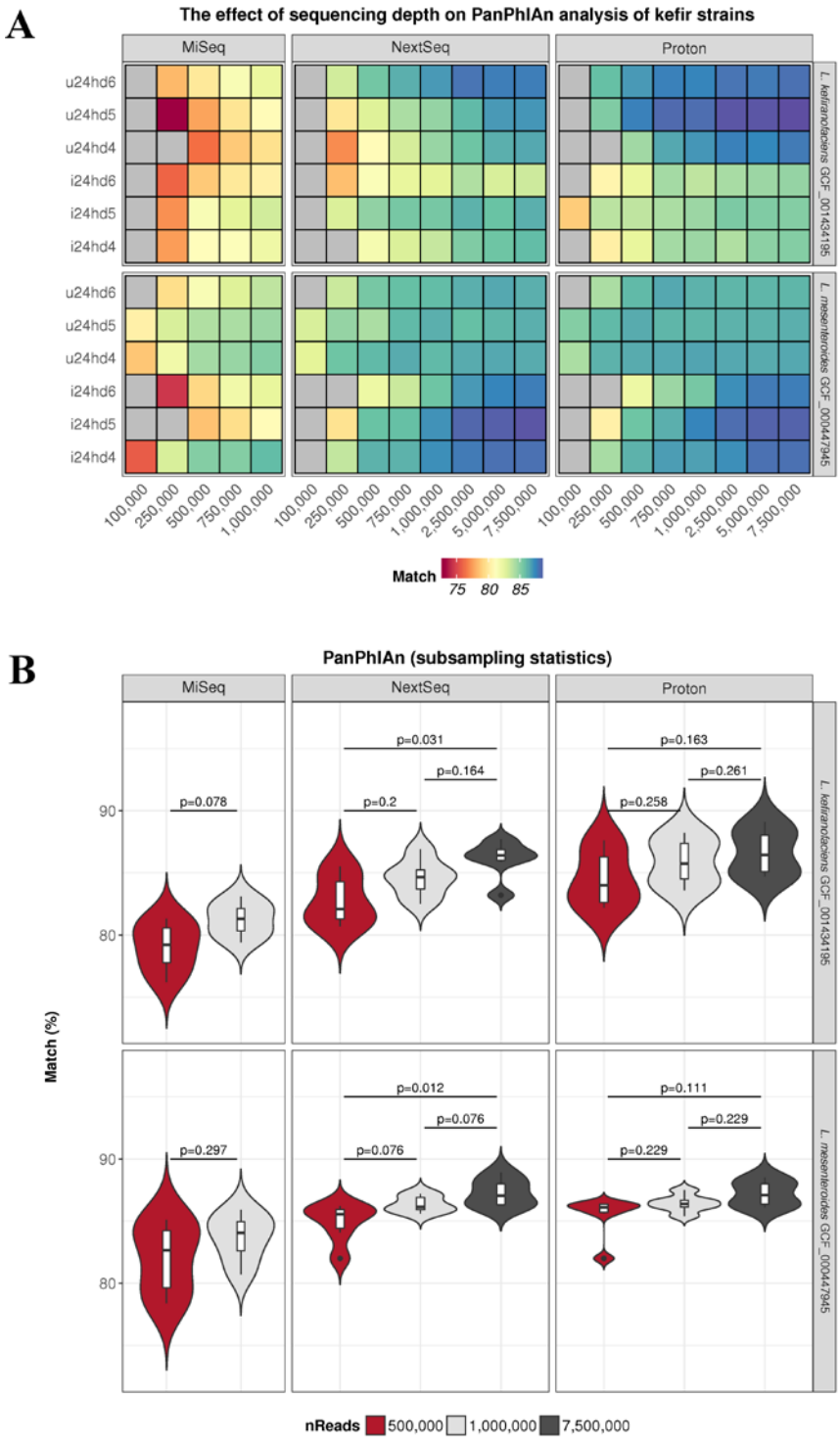


Figure 8: The effect of sequencing depth on PanPhlAn analysis of the two most abundant kefir species, *Lactobacillus kefiranofaciens* and *Leuconostoc mesenteroides.* (A) The predicted percentage similarity of kefir strains relative to their most closely related reference strain, at each sequencing depth. Grey cells indicate that the species was not classified to the strain-level at the specified depth. (B) Statistical differences in the percentage similarity at 100,000, 1,000,000 and 7,500,000 reads per sample.

mesenteroides: p=0.012) (Figure 8B). Overall, our results indicate that the tool's accuracy improves with increased sequencing depth.

The reproducibility of random subsampling improves with increased sequencing depth

The reproducibility of sequence subsampling was assessed by randomly subsampling each kefir sample 10 times at 100,000 reads, 250,000 reads, and 500,000 reads. The subsampled reads were analysed using MetaPhlAn2 and SUPER-FOCUS. For MetaPhlAn2, MDS showed that replicates clustered together at each sequencing depth (Figure S9A). However, the average distance from replicates to their respective centroids significantly decreased with increased sequencing depth for each sequencer (Figure S9B). Additionally, at 500,000 reads, the distance to the centroid was significantly lower for the MiSeq than either the NextSeq or the Proton (Figure S9C). Similarly, for SUPER-FOCUS, MDS showed that replicates clustered together at each sequencing depth (Figure S10A). However, again, the distance to the centroid significantly decreased with increased sequencing depth for each sequencer (Figure S10B). Furthermore, at all sequencing depths, the distance to the centroid was lower for the MiSeq than either the NextSeq or the Proton, and it was also lower for the NextSeq than the Proton (Figure S10C). Overall, our results indicate that random subsampling is consistent but reproducibility does improve with sequencing depth. The MiSeq gave the most consistent results, which is perhaps because it produces longer read lengths than the other two platforms.

Discussion

Currently, there is no consensus as to which next generation sequencing platforms are most suitable for shotgun metagenomics of low complexity microbial communities, such as those in foods. Optimised determination of food microbiota is of considerable relevance to ensuring the safety, quality and health-promoting attributes of foods. Here, we use a variety of bioinformatic tools to benchmark the performances of three high-throughput platforms for shotgun metagenomics of food microbial communities: the Illumina MiSeq, the Illumina NextSeq, and the Ion Proton. Our results highlight a remarkable similarity in the results generated with each platform in terms of compositional, functional, and strain-level analysis. In contrast, several issues with the outputs from species classifiers were identified. Notably, the results of MetaPhlAn2 analysis differed from those of the other species classifiers. We expect that this is because MetaPhlAn2 is based on the alignments with species-specific marker gene sequences, whereas the other methods, which can be categorised as taxanomic binning tools, are based on alignments with whole genome sequences. In fact, we noted that the relative abundances of mock community species, as predicted by all of the species classifiers apart from MetaPhlAn2, correlated to the size of their respective reference genomes. Thus, our results confirm previous observations that these species classifiers are biased by the size of the reference genome (58), in the same way that 16S rRNA gene sequencing is biased by the number of 16S rRNA genes per genome. It is important to be aware of this issue when reporting species abundances. A potential solution to the problem is to normalise relative abundances by genome size. Indeed, this solution has already been suggested elsewhere (58, 59), and we found that normalisation resulted in a more even species distribution. However, this solution is limited by the assumption

that intraspecific strains share the same genome sizes, when, in fact, genome sizes often vary within a species (60). We noted some additional discrepancies between the species classifiers. Specifically, *Corynebacterium casei* was overlooked within the mock community by CLARK or Kraken, even though the species was present in their respective databases. Compositional analysis of the mock community also produced numerous probable false positive species classifications, especially in the case of SLIMM, but most of the false positives were closely related to the actual mock community species and they were present at less than 1% relative abundance. Overall, our results indicated that none of the classifiers are entirely accurate, but we suggest that MetaPhlAn2, and perhaps Kaiju, are the most suitable for compositional analysis of low complexity communities, especially foods, since both tools identified all of the mock community species and they can additionally detect eukaryotic organisms.

Compositional analysis of kefir showed that the choice of sequencing platform did not noticeably affect the results. However, dissimilarity analysis again highlighted marked differences between the outputs generated by the species classifiers. Thus, for compositional analysis, the choice of sequencing platform had less of an influence on results than the choice of species classifier. These observations are consistent with findings from a previous sequencing platform comparison study (34), where the authors demonstrated that gut metagenome samples clustered by species classifier. Such results highlight a need for consistency in bioinformatics methodologies across studies, but the issue is confounded by the increasing availability of different species classifiers. The recently developed method MetaMeta (59), which integrates the results from multiple species classifiers to mitigate the flaws from each individual tool, might partially address this problem.

We did not use MetaMeta here because the default program employs a different combination of species classifiers to that used in our study. Instead, we averaged the predicted taxonomic profiles from each species classifier for every sample, as an alternative solution, and subsequent analysis confirmed that there was no significant dissimilarity between the sequencers. Another possible option for compositional analysis, which we did not explore here, is to use a *de novo* metagenome assembly approach, wherein genomes are binned using tools like CONCOCT (61) or MetaBAT (62), and reads are then mapped against these bins to calculate species abundances. An advantage of such an approach is that it does not rely on a reference database for diversity analysis, and it may also be able to estimate the abundances of potentially novel genomes. However, sequence alignment against a reference database is still necessary to assign taxonomy to the bins, and, additionally, the approach requires a considerably higher sequencing depth than short-read alignment-based methods (63).

Another important aspect of shotgun metagenomics is its ability to characterise the functional potential of metagenomes. Again, the results of functional analysis were generally consistent between all three sequencing platforms, but SUPER-FOCUS did detect significant differences in three functions which were present at greater than 1% relative abundance within the kefir metagenome. Such discrepancies suggest that results generated with different sequencers cannot be reliably compared.

Above, we described a considerable difference in the compositional profiles determined by different species classifiers. Hence, we also compared results from SUPER-FOCUS with those from HUMAnN2, which is an alternative tool for functional analysis of metagenomes. We observed a similarly pronounced disparity in the results generated by these methods. Specifically, 865 level-4 enzyme

commission (EC) categories were detected with both tools, but 749 of these EC categories were differentially abundant between them. Our observation is not unexpected since these pipelines use inherently different approaches, but it does further emphasise that results obtained using different methods cannot be directly compared.

Next, we compared the results of strain-level analysis using PanPhlAn, and we found that all three sequencers correctly identified the analysed strains from the mock community sample. Similarly, the three platforms each indicated that the *L. kefiranofaciens* and *L. mesenteroides* strains detected in the kefir samples were most closely related to *L. kefiranofaciens* GCF_001434195 and *L. mesenteroides* GCF_000447945, respectively. PanPhlAn was significantly less accurate when utilising data generated by the MiSeq compared to either NextSeq or Proton data, suggesting that sequencing depth affected strain-level analysis. We subsequently confirmed this by randomly subsampling kefir sequencing reads which demonstrated that PanPhlAn failed to detect *L. kefiranofaciens* GCF_001434195 or *L. mesenteroides* GCF_000447945 a subset of kefir samples below 500,000 reads per sample using any sequencer. Similarly, and as expected, we observed that sequencing depth significantly improved metagenome assembly completeness. On the other hand, sequencing depth did not have a noticeable effect on compositional or functional analysis of the mock community or kefir, regardless of the choice of sequencer. Indeed, the results of these analyses were almost uniform at sequencing depths ranging from 100,000 reads per sample to 7,500,000 reads per sample, regardless of the choice of species classifier. It is important to note, however, that increased sequencing depth caused a slight, but significant, improvement in the

reproducibility of random subsampling, which suggests that higher coverage offers more reproducible results.

Overall, our findings confirm that the Proton is on par with Illumina sequencers in terms of accuracy, but only a handful of studies have used the Proton for shotgun metagenomics to date (64, 65), even if it is widely used for human exome sequencing. On the basis of these investigations the Proton is a viable option for metagenomic analyses.

To date, most high-throughput sequencing-based studies of microbial communities of food have relied upon 16S rRNA gene sequencing (35). Shotgun metagenomics can, in general, offer higher taxonomic resolution than amplicon sequencing, although the latter approach may be superior for studying poorly microbiologically characterised environments that contain few species for which there are available reference genomes. Shotgun metagenomics can also be used for the direct functional characterisation of metagenomes. Several recent studies have demonstrated the enormous potential for shotgun metagenomic analysis of foods, and indeed, we have previously used the approach to: identify the cause of a pink discoloration defect in Swiss-type cheeses (66), link microbial species with distinct flavours during kefir fermentation (39), and identify pathogenic strains in nunu (67). However, the higher cost of shotgun metagenomics is considered prohibitive for commercial application of the technology by the food industry and, consequently, the approach has been relatively underutilised. This is partially due to a perception that shotgun metagenomics requires considerable sequencing depth per sample. Notably, our results suggest that this is not necessarily true for the low complexity microbial communities present in foods and suggest that 750,000 to 1,000,000 reads per

sample is sufficient for compositional and/or functional analysis of such simple communities.

Conclusion

In conclusion, analysis of low diversity metagenomic DNA representative of food microbial communities highlighted that outputs were consistent across a variety of sequencing platforms at different sequencing depths, but there were clear disparities between the outputs of bioinformatics tools. Thus, the choice of sequencer for shotgun metagenomics can be dictated by logistical factors, like platform availability, budget or sample size, rather than sequencing chemistry. It is hoped that this work will guide researchers, particularly food microbiologists, in designing shotgun metagenomics experiments to exploit the extensive possibilities offered by the approach.

References

1. **Consortium HMP.** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486:**207.
2. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG.** 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences **109:**21390-21395.
3. **Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting L, Ertan H, Johnson J.** 2009. The genomic basis of trophic strategy in marine bacteria. Proceedings of the National Academy of Sciences **106:**15527-15533.
4. **Gilbert JA, Dupont CL.** 2011. Microbial metagenomics: beyond the genome. Annual Review of Marine Science **3:**347-371.
5. **Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL.** 2016. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and Biophysical Research Communications **469:**967-977.
6. **Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R.** 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology **12:**635.
7. **Noecker C, McNally CP, Eng A, Borenstein E.** 2017. High-resolution characterization of the human microbiome. Translational Research **179:**7-23.

8. **Allard G, Ryan FJ, Jeffery IB, Claesson MJ.** 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinformatics **16:**1.

9. **Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML.** 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution **4:**1111-1119.

10. **Lindgreen S, Adair KL, Gardner PP.** 2016. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports **6:**19233.

11. **Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D.** 2015. ConStrains identifies microbial strains in metagenomic datasets. Nature Biotechnology **33:**1045.

12. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

13. **Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N.** 2017. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research **27:**626-638.

14. **Zolfo M, Tett A, Jousson O, Donati C, Segata N.** 2016. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Research**:**gkw837.

15. **Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C.** 2015. Sequencing and beyond: integrating

molecular'omics' for microbial community profiling. Nature Reviews Microbiology **13:**360-372.

16. **Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R.** 2012. Unlocking the potential of metagenomics through replicated experimental design. Nature Biotechnology **30:**513-520.

17. **Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP.** 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics **15:**121-132.

18. **Reuter JA, Spacek DV, Snyder MP.** 2015. High-throughput sequencing technologies. Molecular Cell **58:**586-597.

19. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456:**53-59.

20. **Dubin K, Callahan MK, Ren B, Khanin R, Viale A, Ling L, No D, Gobourne A, Littmann E, Huttenhower C.** 2016. Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. Nature Communications **7:**10391.

21. **Milani C, Ticinesi A, Gerritsen J, Nouvenne A, Lugli GA, Mancabelli L, Turroni F, Duranti S, Mangifesta M, Viappiani A.** 2016. Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. Scientific Reports **6**.

22. **Yergeau E, Michel C, Tremblay J, Niemi A, King TL, Wyglinski J, Lee K, Greer CW.** 2017. Metagenomic survey of the taxonomic and functional

microbial communities of seawater and sea ice from the Canadian Arctic. Scientific Reports **7**.

23. **Deng X, den Bakker HC, Hendriksen RS.** 2016. Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. Annual Review of Food Science and Technology **7:**353-374.

24. **Speth DR, In't Zandt MH, Guerrero-Cruz S, Dutilh BE, Jetten MS.** 2016. Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. Nature Communications **7**.

25. **Ni J, Ramkissoon SH, Xie S, Goel S, Stover DG, Guo H, Luu V, Marco E, Ramkissoon LA, Kang YJ.** 2016. Combination inhibition of PI3K and mTORC1 yields durable remissions in mice bearing orthotopic patient-derived xenografts of HER2-positive breast cancer brain metastases. Nature Medicine **22:**723-726.

26. **Riera M, Navarro R, Ruiz-Nogales S, Méndez P, Burés-Jelstrup A, Corcóstegui B, Pomares E.** 2017. Whole exome sequencing using Ion Proton system enables reliable genetic diagnosis of inherited retinal dystrophies. Scientific Reports **7:**42078.

27. **Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, Zhang L-H, Bhavsar AP, Lee JJ, Drögemöller BI.** 2016. Exome sequencing and the management of neurometabolic disorders. New England Journal of Medicine **374:**2246-2255.

28. **Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M.** 2011. An integrated

semiconductor device enabling non-optical genome sequencing. Nature **475:**348-352.

29. **Ashktorab H, Azimi H, Nickerson ML, Bass S, Varma S, Brim H.** 2016. Targeted Exome Sequencing Outcome Variations of Colorectal Tumors within and across Two Sequencing Platforms. Next Generation, Sequencing & Applications **3**.

30. **Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ.** 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology **30:**434-439.

31. **Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG.** 2014. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Applied and Environmental Microbiology **80:**7583-7591.

32. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13:**341.

33. **Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD.** 2016. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiology **16:**123.

34. **Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, Claesson MJ.** 2016. Comparing Apples and Oranges?: Next Generation

Sequencing and Its Impact on Microbiome Analysis. PLoS One **11:**e0148028.

35. **De Filippis F, Parente E, Ercolini D.** 2017. Metagenomics insights into food fermentations. Microbial Biotechnology **10:**91-102.

36. **Doyle CJ, O'Toole PW, Cotter PD.** 2017. Metagenome-based surveillance and diagnostic approaches to studying the microbial ecology of food production and processing environments. Environmental Microbiology **19**: 4382-4391

37. **Wolfe BE, Button JE, Santarelli M, Dutton RJ.** 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell **158:**422-433.

38. **Pruitt KD, Tatusova T, Maglott DR.** 2006. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research **35:**D61-D65.

39. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.

40. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

41. **Peng Y, Leung HC, Yiu S-M, Chin FY.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28:**1420-1428.

42. **Ounit R, Wanamaker S, Close TJ, Lonardi S.** 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics **16:**1.

43. **Menzel P, Ng KL, Krogh A.** 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications **7**.

44. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15:**R46.

45. **Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N.** 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods **12:**902-903.

46. **Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K.** 2017. SLIMM: species level identification of microorganisms from metagenomes. PeerJ **5:**e3138.

47. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9:**357-359.

48. **Silva GGZ, Green KT, Dutilh BE, Edwards RA.** 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics **32:**354-361.

49. **Buchfink B, Xie C, Huson DH.** 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods **12:**59-60.

50. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B.** 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Computational Biology **8:**e1002358.

51. **Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R.** 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research **33:**5691-5702.

52. **Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U.** 2014. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31:**926-932.

53. **Team RC.** 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.

54. **Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M.** 2007. The vegan package. Community Ecology Package **10:**631-637.

55. **Harrell Jr FE, Harrell Jr MFE.** 2017. Package 'Hmisc'. R Foundation for Statistical Computing.

56. **Wickham H.** 2016. ggplot2: elegant graphics for data analysis. Springer.

57. **Nayfach S, Pollard KS.** 2016. Toward accurate and quantitative comparative metagenomics. Cell **166:**1103-1116.

58. **Nayfach S, Pollard KS.** 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biology **16:**51.

59. **Piro VC, Matschkowski M, Renard BY.** 2017. MetaMeta: Integrating Metagenome Analysis Tools To Improve Taxonomic Profiling. bioRxiv**:**138578.

60. **Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T.** 2015. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics **15:**141-161.

61. **Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C.** 2014. Binning metagenomic contigs by coverage and composition. Nature Methods **11:**1144.

62. **Kang DD, Froula J, Egan R, Wang Z.** 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3:**e1165.

63. **Quince C, Walker AW, Simpson JT, Loman NJ, Segata N.** 2017. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35:**833.

64. **Noyes NR, Yang X, Linke LM, Magnuson RJ, Cook SR, Zaheer R, Yang H, Woerner DR, Geornaras I, McArt JA.** 2016. Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. Scientific Reports **6:**24645.

65. **Kumaresan D, Cross AT, Moreira-Grez B, Kariman K, Nevill P, Stevens J, Allcock RJ, O'Donnell AG, Dixon KW, Whiteley AS.** 2017. Microbial Functional Capacity Is Preserved Within Engineered Soil Formulations Used In Mine Site Restoration. Scientific Reports **7**.

66. **Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, Beresford TP, Ross RP, Fitzgerald GF, McSweeney PLH, Giblin L, Sheehan JJ, Cotter PD.** 2016. Thermus and the Pink Discoloration Defect in Cheese. mSystems **1**.

67. **Walsh AM, Crispie F, Daari K, O'Sullivan O, Martin JC, Arthur CT, Claesson MJ, Scott KP, Cotter PD.** 2017. Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks. Applied and environmental microbiology**:**AEM. 01144-01117.

SUPPLEMENTAL MATERIAL

Compositional analysis

Here, we outline the commands used for each species classifier, in addition to PanPhlAn, and we describe how these parameters deviated from the default settings. Commands are highlighted in grey.

CLARK

```
classify_metagenome.sh -O sample.fasta -R sample.clark.out -m 0
estimate_abundance.sh -F sample.csv -D $DIR_DB -a 0.1 -c 1 -g 0.05
```

Description: The CLARK classification step was run with full mode execution. The CLARK estimate abundances step was run with minAbundance 0.1, minConfidenceScore 1, and minGamma 0.05.

Kaiju

```
kaiju -t $KAIJU_DIR/nodes.dmp -f $KAIJU_DIR/kaiju_db.fmi -i sample.fasta
-o sample.kaiju.out -z 10 –m 33 -x -v
kaijuReport -u -m 0.1 -t $KAIJU_DIR/nodes.dmp -n $KAIJU_DIR/names.dmp
-r species -i sample.kaiju.out -o sample.species.summary
```

Description: The Kaiju classification step was run using the SEG low complexity filter, and the minimum match length was set to 22. Reads were mapped against the RefSeq database. The Kaiju report step was run using minAbundance 0.1. Only classified reads were reported.

Kraken

kraken --threads 10 --preload --db $KRAKEN_DIR/krakken_db sample.fasta > sample.kraken.out

kraken-filter --db $KRAKEN_DIR/krakken_db --threshold 0.5 sample.kraken.out > sample.kraken.filtered

kraken-mpa-report --db $KRAKEN_DIR/krakken_db sample.kraken.filtered > sample.kraken.mpa

Description: Kraken results were filtered using a threshold set to 0.5 to remove low confidence classifications.

MetaPhlAn2

MetaPhlAn2 was run using default parameters (https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2).

SLIMM

bowtie2 -x $SLIMM_DB/AB_5K_indexed_ref_genomes_bowtie2/AB_5K -U sample.fastq | samtools view -bSF4 - > sample.slimm.mapped.bam

slimm -m $SLIMM_DB/slimmDB_5K sample.slimm.mapped.bam

Description: Bowtie 2 (1) was used to trimmed fastq reads against the slimmDB_5K reference database.

PanPhlAn

panphlan_map.py -c $pangenome_bowtie2_index -i sample.fasta -o map_results

panphlan_profile.py -c $pangenome_bowtie2_index -i map_results --add_strains --min_coverage 1 --left_max 1.70 --right_min 0.30 --o_dna result_gene_presence_absence.csv --strain_hit_genes_perc percent_match.txt

Description: The PanPhlAn profiling step was run with a –min_coverage set to 1, --left_max set to 1.7, and –right_min set to 0.3. These parameters increase the tool's sensitivity.

REFERENCES

1. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nature methods **9:**357-359.

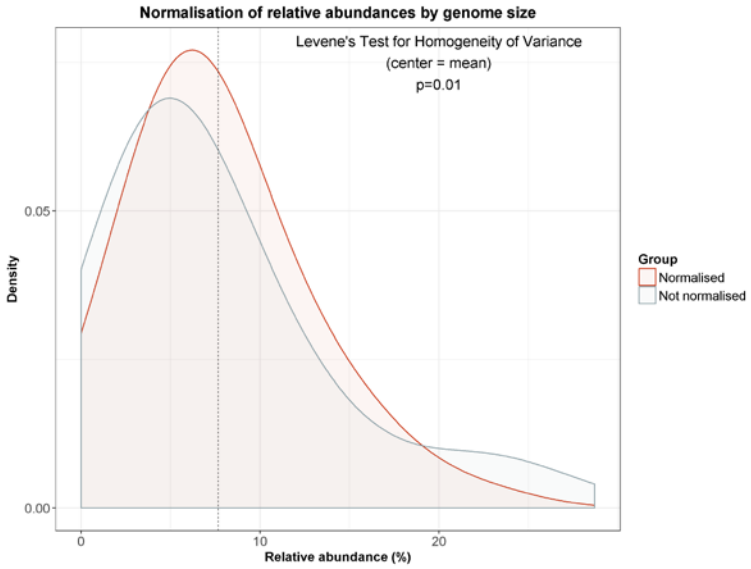


Figure S1: The effect of normalising predicted relative abundances by reference genome size. The histogram shows the distribution of the relative abundances of the mock community species, before and after normalisation. The results are averaged across sequencers and metagenome binning tools (i.e. CLARK, Kaiju, Kraken, and SLIMM).

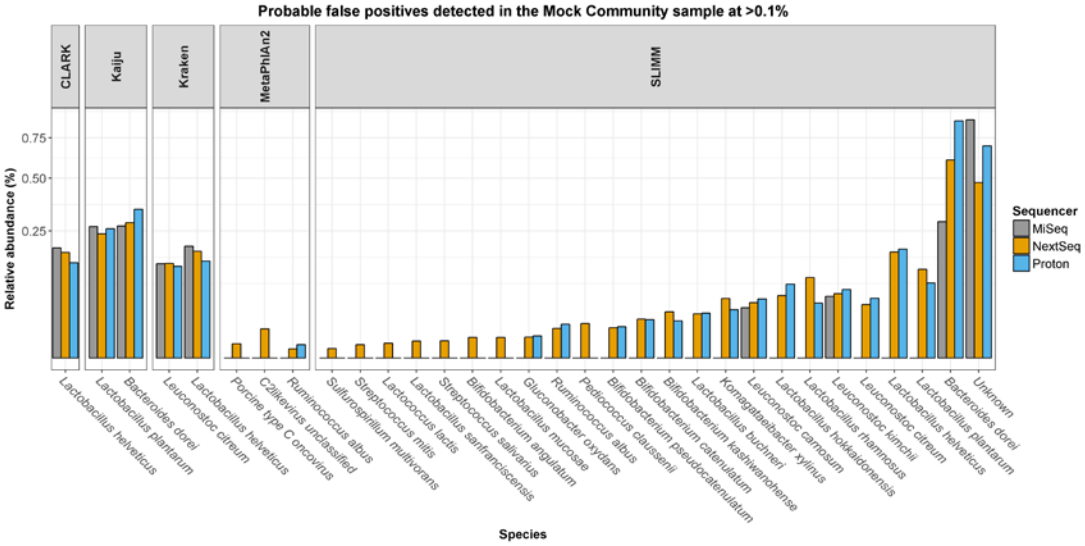


Figure S2: False positives detected using each species classifier with the total number of reads from each sequencer.

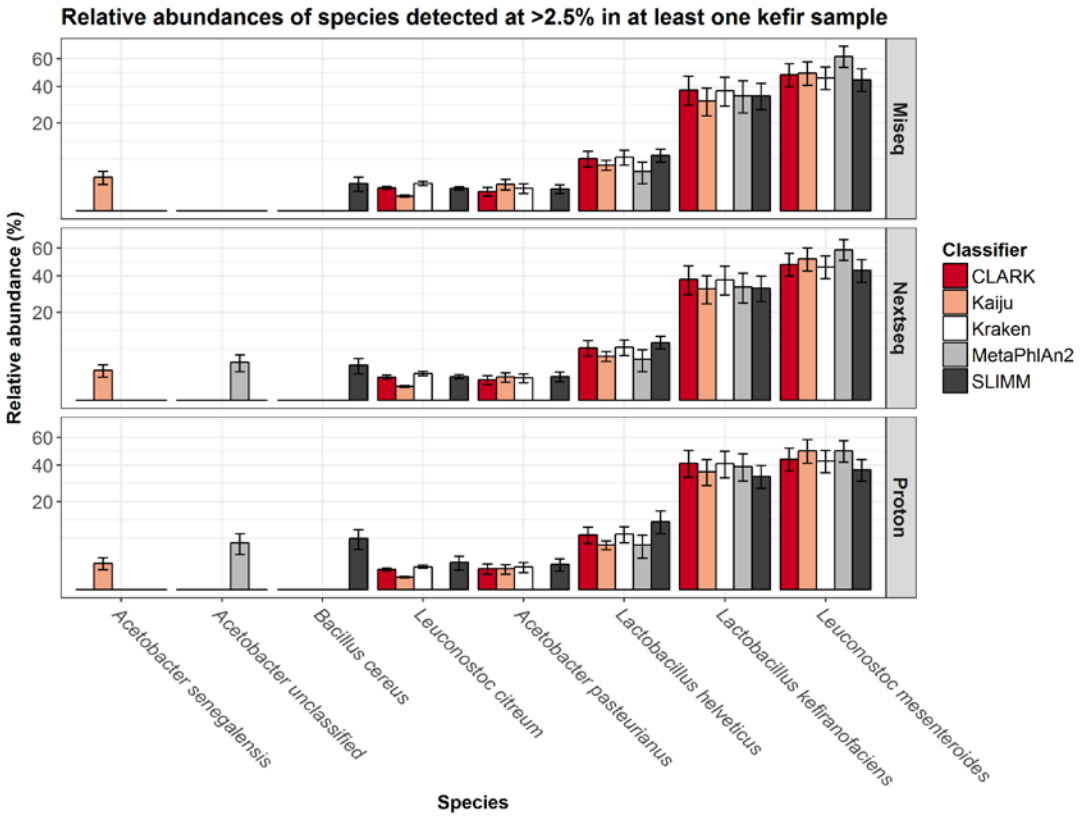


Figure S3: Species detected ≥2.5% relative abundance in kefir samples using each species-classifier with the total number of reads from each sequencer.

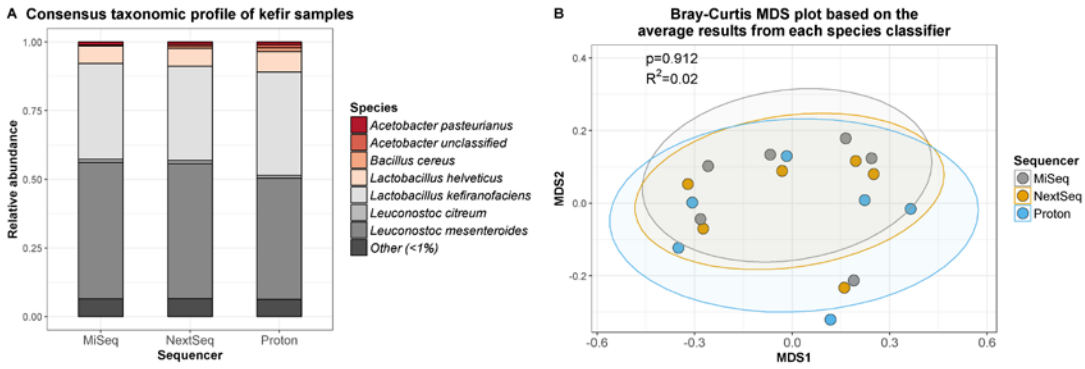


Figure S4: (A) The consensus taxonomic profile of kefir samples, as predicted by averaging the results from each species classifier. (B) Dissimilarity plot based on the average results from each species classifier.

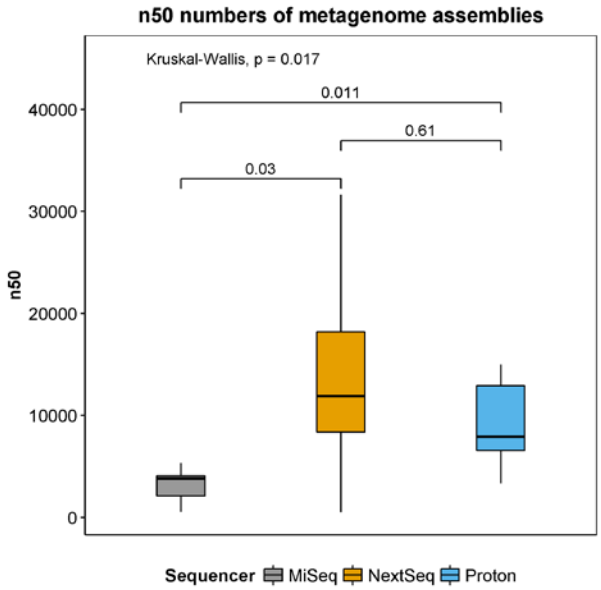


Figure S5: n50 number of metagenome assemblies which were assembled using the total number of reads from each sequencer.

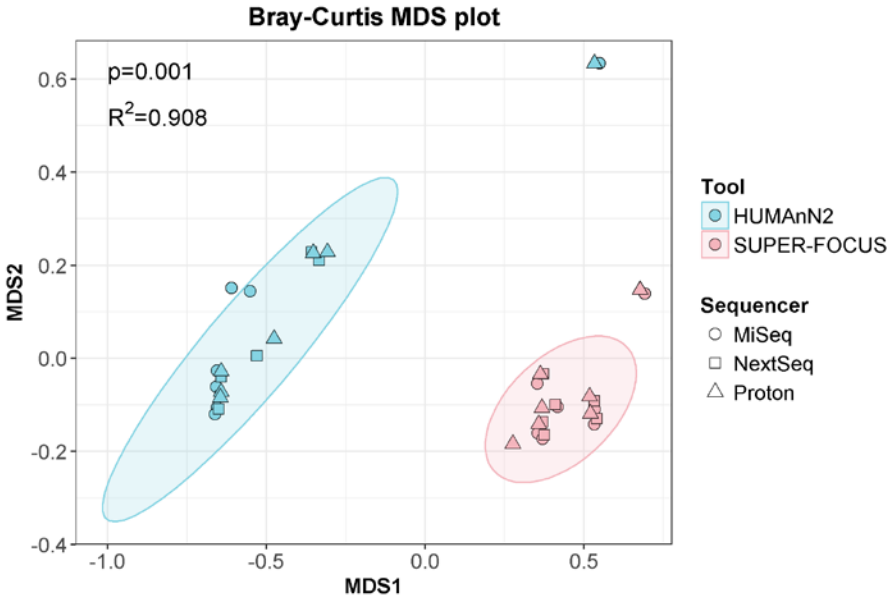


Figure S6: Dissimilarity plot based on the relative abundances of the 865 level-4 enzyme commission (EC) categories which were detected by both HUMAnN2 and SUPER-FOCUS.

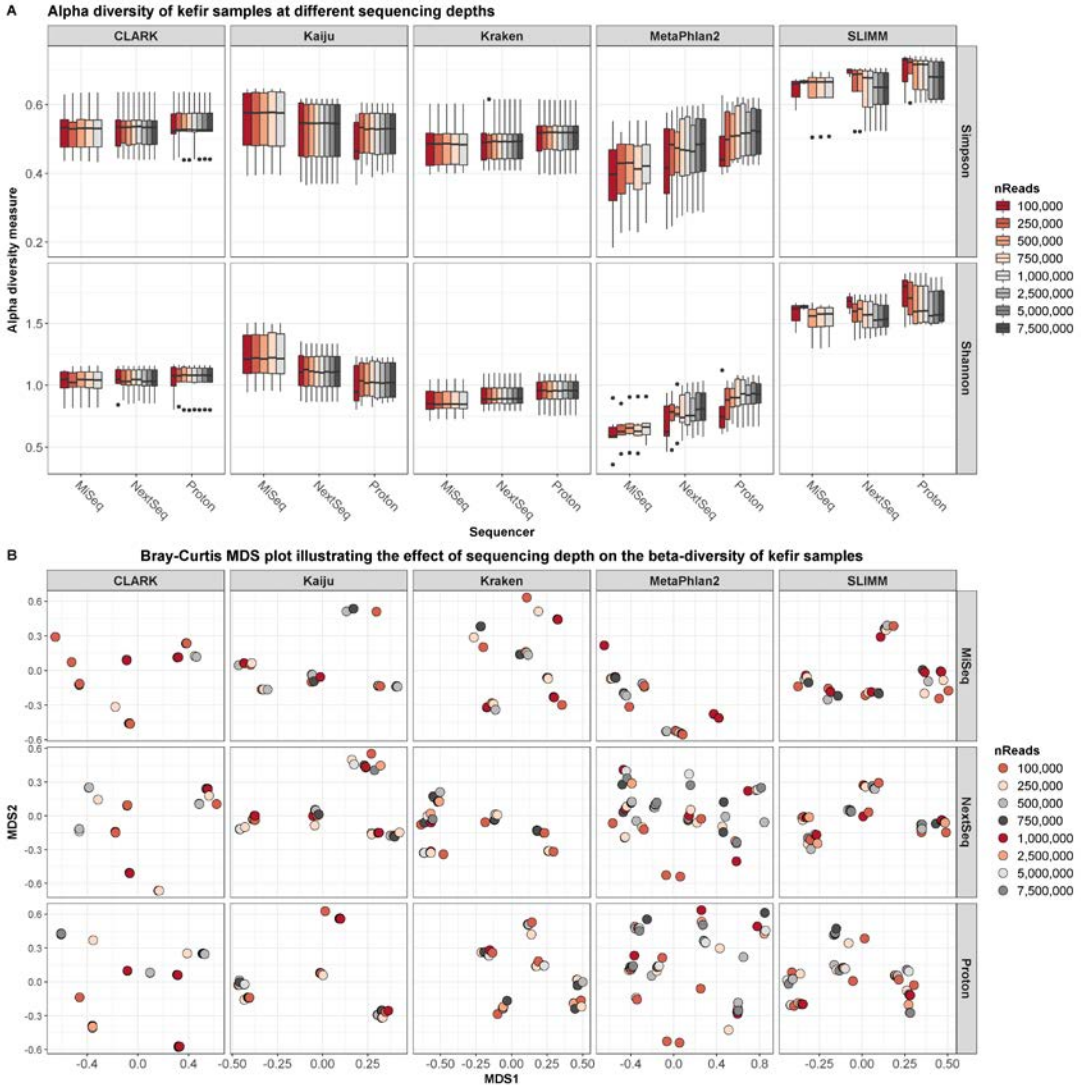


Figure S7: The effect of subsampling on the predicted diversity of kefir samples. (A) The alpha-diversity of kefir samples at different sequencing depths on each sequencer. (B) Dissimilarity plot based on the relative abundances of the compositional analysis of subsampled kefir reads from each sequencer.

Figure S8: SUPER-FOCUS level 2 subsystems which were significantly altered at different sequencing depths.

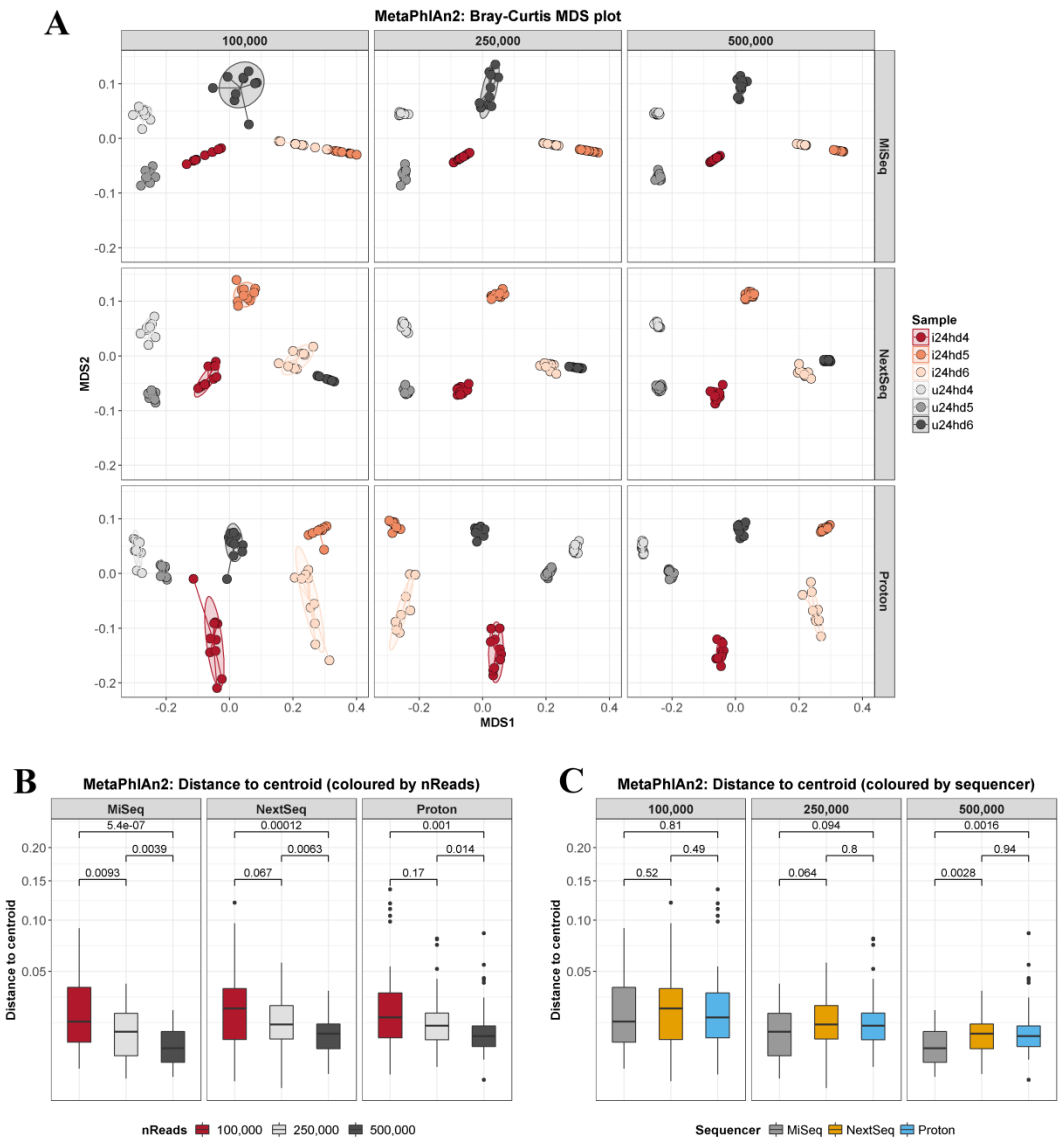


Figure S9: Consistency in the MetaPhlAn2 profiles of randomly subsampled replicates from the same samples. (A) MDS plot (facetted by number of reads) where replicates (coloured by sample) are connected to their respective centroids. (B) The average distance of replicates to their respective centroids at each sequencing depth. (C) The average distance of replicates to their respective centroids for each sequencer.

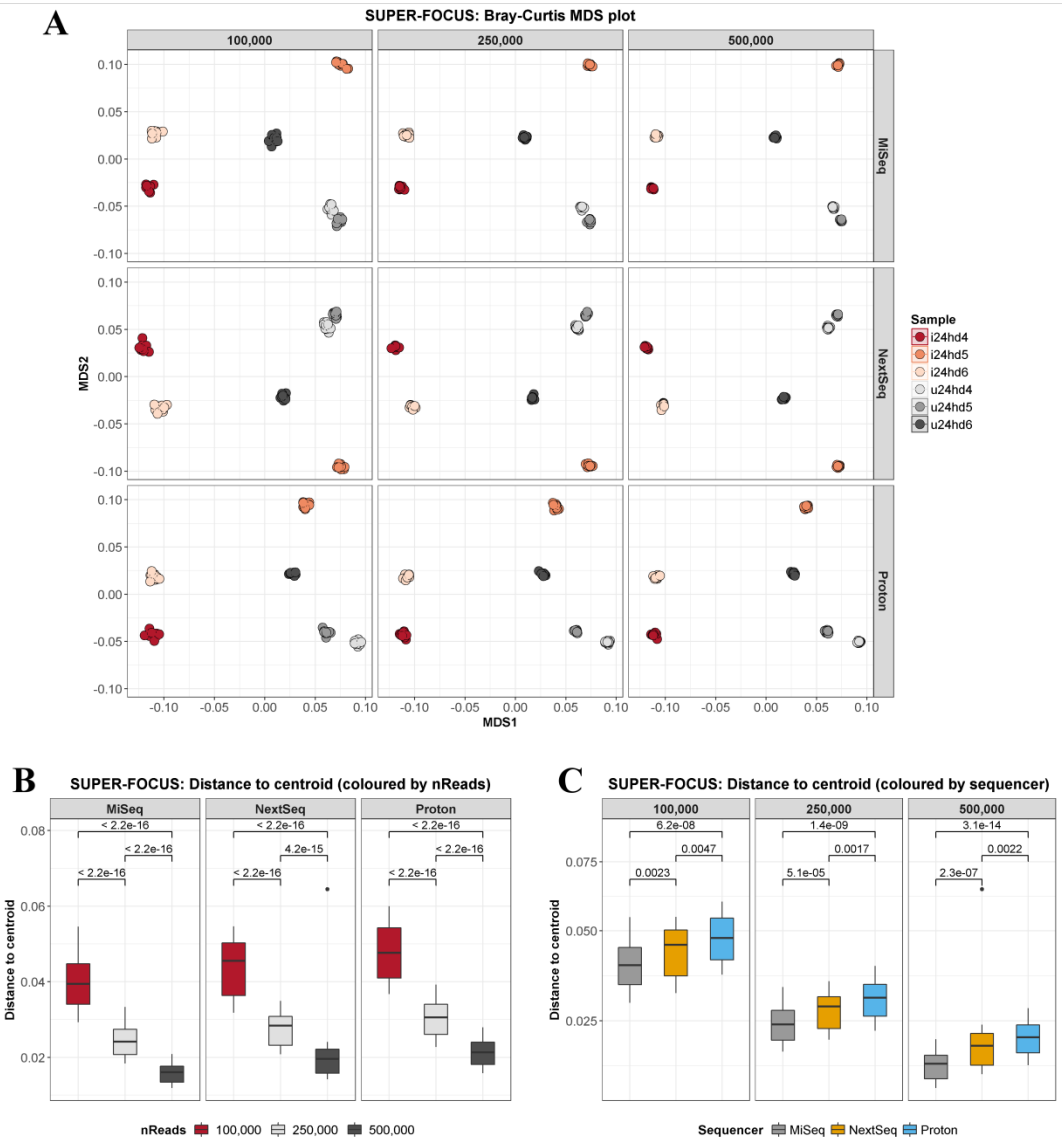


Figure S10: Consistency in the SUPER-FOCUS profiles of randomly subsampled replicates of the same samples. (A) MDS plot (facetted by number of reads) where replicates (coloured by sample) are connected to their respective centroids. (B) The average distance of replicates to their respective centroids at each sequencing depth. (C) The average distance of replicates to their respective centroids for each sequencer.

Table S1: Statistical differences in the alpha diversity of kefir samples between the three sequencers.

Classifier	Index	p.MiSeq vs NextSeq	p.MiSeq vs Proton	p.NextSeq vs Proton
CLARK	Shannon	0.873	0.873	0.873
CLARK	Simpson	1	1	1
Kaiju	Shannon	0.522	0.522	0.522
Kaiju	Simpson	0.635	0.635	0.873
Kraken	Shannon	0.631	0.631	0.631
Kraken	Simpson	0.873	0.873	0.873
MetaPhlAn2	Shannon	0.164	0.075	0.423
MetaPhlAn2	Simpson	0.505	0.235	0.522
SLIMM	Shannon	0.635	0.635	0.749
SLIMM	Simpson	0.337	0.3	0.3

Table S2: Statistical differences in the alpha diversity of kefir samples between species classifiers.

Sequencer	Index	p.CLARK vs Kaiju	p.CLARK vs Kraken	p.CLARK vs MetaPhlAn2	p.CLARK vs SLIMM	p.Kaiju vs Kraken	p.Kaiju vs MetaPhlAn2	p.Kaiju vs SLIMM	p.Kraken vs MetaPhlAn2	p.Kraken vs SLIMM	p.MetaPhlAn2 vs SLIMM
MiSeq	Shannon	0.098	0.053	0.053	0.016	0.522	0.016	0.166	0.016	0.033	0.016
MiSeq	Simpson	0.421	0.421	0.58	0.273	0.873	0.4	0.421	0.4	0.364	0.25
NextSeq	Shannon	0.098	0.053	0.053	0.016	0.522	0.016	0.166	0.016	0.033	0.016
NextSeq	Simpson	0.421	0.421	0.58	0.273	0.873	0.4	0.421	0.4	0.364	0.25
Proton	Shannon	0.47	0.098	0.098	0.02	0.873	0.075	0.098	0.054	0.062	0.02
Proton	Simpson	0.701	0.481	0.701	0.082	0.749	0.481	0.137	0.481	0.083	0.065

Table S3: Statistical differences in the predicted species relative abundances between classifiers.

Sequencer	Species	p.overall	p.CLARK vs Kaiju	p.CLARK vs Kraken	p.CLARK vs MetaPhlAn2	p.CLARK vs SLIMM	p.Kaiju vs Kraken	p.Kaiju vs MetaPhlAn2	p.Kaiju vs SLIMM	p.Kraken vs MetaPhlAn2	p.Kraken vs SLIMM	p.MetaPhlAn2 vs SLIMM
MiSeq	Acetobacter pasteurianus	0.006	0.47	0.47	0.018	0.47	0.47	0.007	0.47	0.007	1	0.007
MiSeq	Acetobacter senegalensis	0.001	0.002	NA	NA	NA	0.002	0.002	0.002	NA	NA	NA
MiSeq	Acetobacter sp. SLV 7	0.001	0.007	NA	NA	NA	0.007	0.007	0.007	NA	NA	NA
MiSeq	Acetobacter unclassified	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MiSeq	Bacillus cereus	0.008	NA	0.14	NA	0.052	0.14	NA	0.052	0.14	0.14	0.052
MiSeq	Gluconacetobacter diazotrophicus	0.002	0.065	0.05	0.353	0.077	0.328	0.025	0.286	0.01	0.423	0.01
MiSeq	Gluconobacter oxydans	0.003	0.082	0.04	0.453	0.04	0.58	0.025	0.58	0.01	0.749	0.01
MiSeq	Kozakia baliensis	0.001	0.022	NA	NA	NA	0.022	0.022	0.022	NA	NA	NA
MiSeq	Lactobacillus acidophilus	0.001	0.003	0.003	NA	0.003	0.004	0.003	0.004	0.003	0.749	0.003
MiSeq	Lactobacillus buchneri	0.002	0.088	0.005	NA	0.005	1	0.088	1	0.005	0.337	0.005
MiSeq	Lactobacillus helveticus	0.391	0.631	0.58	0.489	0.58	0.58	0.58	0.58	0.489	0.58	0.489
MiSeq	Lactobacillus kefiranofaciens	0.913	0.832	0.832	0.832	0.832	0.832	0.873	0.832	0.832	0.832	0.832
MiSeq	Lactobacillus plantarum	0.001	0.039	0.051	0.317	0.007	0.036	0.007	0.007	0.007	0.007	0.007
MiSeq	Lactobacillus sanfranciscensis	0.005	0.749	0.701	0.005	0.605	0.701	0.005	0.525	0.005	0.605	0.005
MiSeq	Lactococcus lactis	0.011	0.529	0.529	0.146	0.577	0.529	0.007	0.631	0.007	0.525	0.007
MiSeq	Leuconostoc carnosum	0.001	0.053	0.007	0.155	0.007	0.068	0.007	0.053	0.007	0.423	0.007
MiSeq	Leuconostoc citreum	0.001	0.006	0.121	0.005	0.749	0.006	0.005	0.006	0.005	0.121	0.005
MiSeq	Leuconostoc gelidum	0.003	0.178	0.37	0.117	0.747	0.25	0.007	0.062	0.007	0.25	0.007
MiSeq	Leuconostoc kimchii	0.003	0.25	0.155	0.056	0.181	0.58	0.007	0.25	0.007	0.749	0.007
MiSeq	Leuconostoc mesenteroides	0.456	0.873	0.701	0.5	0.701	0.701	0.5	0.701	0.5	0.701	0.5
NextSeq	Acetobacter pasteurianus	0.008	0.701	0.701	0.018	0.701	0.701	0.007	0.873	0.007	0.701	0.007
NextSeq	Acetobacter senegalensis	0.001	0.002	NA	NA	NA	0.002	0.002	0.002	NA	NA	NA
NextSeq	Acetobacter sp. SLV 7	0.001	0.007	NA	NA	NA	0.007	0.007	0.007	NA	NA	NA
NextSeq	Acetobacter unclassified	0.001	NA	NA	0.002	NA	NA	0.002	NA	0.002	NA	0.002
NextSeq	Bacillus cereus	0.006	NA	0.068	NA	0.052	0.068	NA	0.052	0.068	0.135	0.052
NextSeq	Gluconacetobacter diazotrophicus	0.004	0.082	0.077	0.374	0.082	0.374	0.025	0.374	0.01	0.423	0.01
NextSeq	Gluconobacter oxydans	0.006	0.18	0.044	0.453	0.044	0.873	0.044	0.873	0.01	0.873	0.01
NextSeq	Kozakia baliensis	0.012	0.059	NA	NA	NA	0.059	0.059	0.059	NA	NA	NA
NextSeq	Lactobacillus acidophilus	0.001	0.003	0.003	NA	0.003	0.004	0.003	0.004	0.003	0.016	0.003
NextSeq	Lactobacillus buchneri	0.002	0.088	0.005	NA	0.005	1	0.088	1	0.005	0.544	0.005
NextSeq	Lactobacillus helveticus	0.35	0.631	0.631	0.5	0.525	0.605	0.631	0.5	0.5	0.561	0.5
NextSeq	Lactobacillus kefiranofaciens	0.893	0.87	0.873	0.87	0.87	0.87	0.873	0.873	0.87	0.87	0.873
NextSeq	Lactobacillus plantarum	0.001	0.191	0.066	0.353	0.008	0.63	0.037	0.008	0.008	0.008	0.008
NextSeq	Lactobacillus sanfranciscensis	0.003	0.374	0.631	0.005	0.374	0.374	0.005	0.299	0.005	0.374	0.005
NextSeq	Lactococcus lactis	0.036	0.629	0.421	0.421	0.421	0.421	0.03	0.421	0.03	0.58	0.03
NextSeq	Leuconostoc carnosum	0.001	0.129	0.024	0.155	0.014	0.053	0.007	0.021	0.007	0.262	0.007
NextSeq	Leuconostoc citreum	0.001	0.006	0.121	0.005	0.522	0.006	0.005	0.006	0.005	0.121	0.005
NextSeq	Leuconostoc gelidum	0.006	0.328	0.328	0.146	0.747	0.328	0.007	0.219	0.007	0.374	0.007
NextSeq	Leuconostoc kimchii	0.002	0.47	0.156	0.018	0.187	0.156	0.007	0.109	0.007	0.873	0.007
NextSeq	Leuconostoc mesenteroides	0.623	0.631	0.631	0.631	0.631	0.631	0.631	0.631	0.631	0.631	0.631
Proton	Acetobacter pasteurianus	0.015	0.873	0.789	0.018	0.789	0.789	0.01	0.789	0.01	0.832	0.018
Proton	Acetobacter senegalensis	0.001	0.002	NA	NA	NA	0.002	0.002	0.002	NA	NA	NA
Proton	Acetobacter sp. SLV 7	0.012	0.059	NA	NA	NA	0.059	0.059	0.059	NA	NA	NA
Proton	Acetobacter unclassified	0.001	NA	NA	0.002	NA	NA	0.002	NA	0.002	NA	0.002
Proton	Bacillus cereus	0.001	NA	0.009	NA	0.009	0.009	NA	0.009	0.009	0.045	0.009
Proton	Gluconacetobacter diazotrophicus	0.017	0.359	0.067	0.421	0.209	0.83	0.146	0.935	0.021	0.421	0.037
Proton	Gluconobacter oxydans	0.029	0.658	0.258	0.279	0.307	0.37	0.195	0.579	0.021	0.873	0.037
Proton	Kozakia baliensis	0.012	0.059	NA	NA	NA	0.059	0.059	0.059	NA	NA	NA
Proton	Lactobacillus acidophilus	0.001	0.005	0.005	NA	0.008	0.006	0.005	0.006	0.005	0.055	0.008
Proton	Lactobacillus buchneri	0.02	0.255	0.103	0.453	0.145	1	0.145	0.896	0.021	0.789	0.037
Proton	Lactobacillus helveticus	0.332	0.701	0.701	0.437	0.437	0.605	0.749	0.437	0.437	0.437	0.437
Proton	Lactobacillus kefiranofaciens	0.928	0.873	0.873	0.873	0.873	0.873	0.873	0.873	0.873	0.873	0.873
Proton	Lactobacillus plantarum	0.002	0.131	0.077	0.317	0.033	0.166	0.025	0.077	0.021	0.078	0.025
Proton	Lactobacillus sanfranciscensis	0.001	0.47	0.749	0.007	0.078	0.47	0.007	0.05	0.007	0.078	0.018
Proton	Lactococcus lactis	0.164	0.684	0.58	0.58	0.58	0.374	0.323	0.58	0.225	0.58	0.323
Proton	Leuconostoc carnosum	0.003	1	0.098	0.098	0.152	0.041	0.01	0.112	0.01	0.97	0.025
Proton	Leuconostoc citreum	0.001	0.006	0.25	0.005	0.749	0.006	0.005	0.006	0.005	0.47	0.005
Proton	Leuconostoc gelidum	0.007	0.281	0.324	0.146	0.935	0.935	0.01	0.156	0.01	0.182	0.025
Proton	Leuconostoc kimchii	0.008	1	0.332	0.056	0.525	0.075	0.01	0.481	0.01	0.701	0.025
Proton	Leuconostoc mesenteroides	0.598	0.529	0.58	0.529	0.529	0.529	0.873	0.529	0.529	0.529	0.529

Table S4: Statistical differences in alpha diversity at different sequencing depths.

Sequencer	Classifier	Index	100,000 reads versus 1,000,000 reads	100,000 reads versus 7,500,000 reads	1,000,000 reads versus 7,500,000 reads
MiSeq	CLARK	Shannon	1	NA	NA
		Simpson	0.873	NA	NA
NextSeq	CLARK	Shannon	1	1	1
		Simpson	1	1	1
Proton	CLARK	Shannon	1	1	1
		Simpson	0.873	0.873	0.873
MiSeq	Kaiju	Shannon	1	NA	NA
		Simpson	1	NA	NA
NextSeq	Kaiju	Shannon	1	1	1
		Simpson	0.873	0.873	0.873
Proton	Kaiju	Shannon	1	1	1
		Simpson	0.873	0.873	0.873
MiSeq	Kraken	Shannon	0.749	NA	NA
		Simpson	1	NA	NA
NextSeq	Kraken	Shannon	1	1	1
		Simpson	0.873	0.873	0.873
Proton	Kraken	Shannon	1	1	1
		Simpson	0.946	0.946	1
MiSeq	MetaPhlAn2	Shannon	0.423	NA	NA
		Simpson	0.631	NA	NA
NextSeq	MetaPhlAn2	Shannon	0.3	0.3	0.873
		Simpson	0.783	0.783	0.873
Proton	MetaPhlAn2	Shannon	0.224	0.224	0.873
		Simpson	0.635	0.635	0.873
MiSeq	SLIMM	Shannon	0.749	NA	NA
		Simpson	0.749	NA	NA
NextSeq	SLIMM	Shannon	0.873	0.873	0.873
		Simpson	0.749	0.749	0.749
Proton	SLIMM	Shannon	1	1	1
		Simpson	0.631	0.631	0.631

Chapter 7

The traditional fermented dairy beverage kefir modulates the gut microbiome and behaviours in mice

Manuscript in preparation

Authors: Aaron M. Walsh, Marcel van de Wouw, Fiona Crispie, Lucas van Leuven, Laura Finnegan, Joshua M. Lyte, Rubén Miranda-Hevia, Marcus Böhme, Trudy Quirke, Gerard Clarke, Timothy G. Dinan, Marcus J. Claesson, John F. Cryan, and Paul D. Cotter.

Contributions:

- **Candidate** performed DNA extractions and computational analysis
- **MvdW** performed animal handling, behavioural analysis, and flow cytometry
- **LvL** assisted with behavioural testing
- **LF**, **RMH**, and **TQ** assisted with DNA extractions and library preparations
- **JML** and **GC** performed HPLC analysis
- **MB** assisted with flow cytometry
- **FC**, **TGD**, **MJC**, **JFC,** and **PDC** supervised the study

ABSTRACT

It is increasingly understood that the gut microbiota can influence behaviour through what has been coined the microbiota-gut-brain axis. In addition, administration of bacterial strains exerting a positive effect on the host (probiotics) can improve mood and have even been termed psychobiotics. Interestingly, this term of psychobiotics has recently been extended to prebiotics. Recent evidence also suggests that fermented foods, which frequently contain probiotics, may affect mood. Here, we investigated if the traditional fermented dairy beverage kefir modulates the microbiota-gut-brain axis in mice. Two distinct kefirs (UK4 and Fr1) or milk control were administered to male adult mice and behaviour was assessed. The kefir UK4 significantly decreased repetitive behaviour and induced a trend towards decreased depressive-like behaviour. Similarly, the kefir Fr1 significantly increased reward-seeking behaviour. Additionally, shotgun metagenomics revealed that both kefirs altered the microbiome along the gastrointestinal tract in the mice. Notably, strain-level analysis indicated that kefir ingestion increased the relative abundances of bacteria containing genes for gamma-aminobutyric acid (GABA) production along with tryptophan biosynthesis. Deficiencies in GABA and the tryptophan derivative serotonin have been linked to anxiety and depression. Thus, our findings show that kefir is able to modulate the microbiota-gut-brain axis and modify mood, potentially by increasing the capacity for the gut microbiome to synthesise neurotransmitters and/or their precursors.

INTRODUCTION

Increasing evidence suggests that the gastrointestinal microbiota can influence host behaviour via bi-directional communication along the gut-brain axis (1). Consequently, the gut microbiota might be a taget for treating disorders such as anxiety or depression (2). Treatment with probiotics, which are live microorganisms that confer health benefits (3), represents one strategy with which to manipulate the gut microbiota (4), and it has been established that some probiotics, also termed 'psychobiotics', can improve mood (5). Additionally, recent data indicates that some prebiotics may be classified as psychobiotics (6). It is also becoming apparent that fermented foods might benefit conditions such as social anxiety (7) or gestational depression (8). Notably, a fermented milk product, which was produced using known probiotics, has been demonstrated to modulate brain activity in healthy women (9). Such findings merit investigation into the mechanisms by which fermented foods might affect the gut-brain axis.

Kefir is a traditional fermented milk beverage that is produced by adding a kefir grain to milk, which is then incubated at room temperature for approximately 24 hours. The kefir grains are exopolysaccharide matrices with what is frequently referred to as 'cauliflower-like' appearance harbouring symbiotic microbial communities, including bacteria and yeasts, which together are responsible for fermentation. The word kefir is derived from the Turkish *keyif*, which translates as "good feeling" (10). Indeed, numerous health benefits have been ascribed to kefir (11) and, consequently, it is frequently described as a natural probiotic beverage (12). It is increasingly understood that kefir microbes are at least partially responsible for these effects (11, 13). Notably, several studies have reported that kefir reduces inflammation in animal models (14-16), while amplicon sequencing

has also revealed that kefir can alter the gut microbiota in mice (17, 18). One of the ways in which the microbiota is able to influence the brain it through modulation of the immune system (19), and, therefore, it is conceivable that kefir might influence behaviour through the gut-brain axis. Intriguingly, a 2014 study showed that kefir reduced nicotine cessation-induced anxiety- and depressive-like behaviour, as well as impairments in long-term spatial learning, in rats (20), but its impact on the gut microbiota was not examined in that study.

Shotgun metagenomics is a powerful tool for characterising the gut microbiota (21), but the approach has rarely been employed to study the microbiome in the context of the gut-brain axis (22), and it has not yet been utilised to characterise the ways in which kefir alters the gut metagenome. Instead, most studies have relied on amplicon sequencing, which typically only offers genus-level information on the composition of microbiota (23). Contrastingly, shotgun metagenomics, which involves determining the entire microbial genetic content within environmental samples, including intestinal samples, yields insights into the functional potential, in addition to the species-level composition, of microbiota (24). Furthermore, several tools, such as PanPhlAn (25) or StrainPhlAn (26), have recently been released which enable strain-level analysis from shotgun metagenomics data. In the present study we employ shotgun metagenomics in parallel with behavioural analysis to investigate the effects of two traditionally prepared kefirs, relative to unfermented milk, on the intestinal microbiota and behaviour of mice. Specifically, shotgun metagenomics was performed to compare the species- and strain-level microbial composition and the functional potential of the microbiome in the ileum, cecum, and faeces of each treatment group. Our results indicate that kefir ameliorates anxious or depressive-like behaviours in mice, while simultaneously increasing the abundance of bacteria

which contain genes associated neurotransmitter production. Thus, we present strong evidence that a traditional fermented food modulates the gut-brain axis.

METHODS

Animals

This study used male C57Bl/6j mice (8 weeks of age on arrival; Envigo, UK; n = 12/group, n = 48 in total). Animals were housed in groups of 4. Food and drinking water were provided *ad libitum* throughout the study. The holding room had a temperature of 21 ± 1 °C and humidity of 55 ± 10% with a 12-hour light/dark cycle (lights on at 7:00 am). Bodyweight was monitored on a weekly basis. Experiments were conducted in accordance with the European Directive 86/609/EEC and the Recommendation 2007/526/65/EC and approved by the Animal Experimentation Ethics Committee of University College Cork. All efforts were made to reduce the number of animals used and to minimise the suffering of these animals.

Experimental timeline and behavioural testing

Animals were habituated for one week prior to the onset of daily kefir administration by oral gavage. After three weeks of treatment, animals were assessed for various their behavioural phenotype using various tests, which were formed in order of least stressful to most stressful to reduce the likelihood of prior behavioural tests influencing subsequent ones (Figure 1). In addition, there was a minimum of 36-hours between tests. The order of testing was as follows: 1) Marble burying test, 2) 3-Chamber social interaction test, 3) Elevated plus maze, 4) Open field test, 5) Tail-

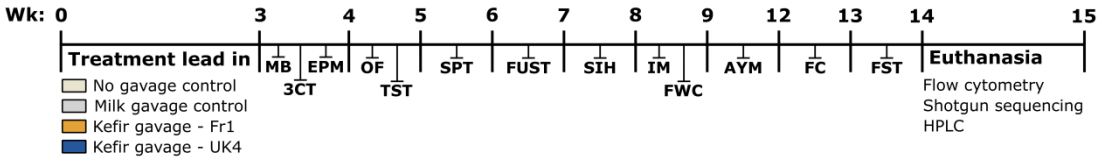


Figure 1: Experimental design. After one week of treatment lead-in, animals were assessed for their behavioural phenotype. Treatment groups consisted of: 1) No gavage control, 2) Milk gavage control, 3) Kefir gavage – Fr1, and 4) Kefir gavage – UK4 (n = 12/group). The order of behavioural tests was as following; Week 4: Marble burying test (MB), 3-Chamber social interaction test (3CT) and Elevate plus maze (EPM); Week 5: Open field test (OF) and Tail suspension test (TST); Week 6: Saccharin preference test (SPT); Week 7: Female urine sniffing test (FUST); Week 8: Stress-induced hyperthermia test (SIH); Week 9: Intestinal motility test (IM) and Faecal water content assessment (FWC): Week 9-12: Appetitive Y-maze; Week 13: Fear conditioning; Week 14: Forced swim test; Week 15: Euthanasia. Postmortem, the immune system was assessed by flow cytometry, Ileal, caecal and faecal microbiota composition and function was investigated by shotgun sequencing, and ileum and colonic serotonergic levels were quantified by high-performance liquid chromatography (HPLC).

suspension test, 6) Saccharin preference test, 7) Female urine sniffing test, 8) Stress-induced hyperthermia test, 9) In testinal motility test, 10) Assessment of faecal water content and weight, 11) Appetitive Y-maze, 12) Fear conditioning, 13) Forced swim test. At the end of the study, body composition (i.e. percentage lean, fat and fluid mass) was assessed (Minispec mq 7.5), after which animals were immediately sacrificed by decapitation.

Kefir culturing and administration

Kefir grains were cultured in whole milk (2% w/v) at 25 °C and milk was renewed every 24 hours using a sterile Buchner funnel and sterile Duran bottle. Grains were rinsed with deionised water prior to the renewal of milk. The fermented milks (i.e. kefirs) collected after the culturing, or milk control, were administered to the mice within one hour by oral gavage (0.2 mL). Daily kefir administration was performed after the behavioural test, if one was performed that day, between 4 pm and 7 pm. To analyse the kefir microbiota over time, aliquots from the kefir administered to the mice were taken on a weekly basis and stored at −80 °C for later analysis.

Marble burying test

Mice were tested for repetitive and anxiety-like behaviour with the marble burying test, which was conducted as previously described (6). Animals were individually placed in a novel Plexiglas cage (35 × 28 × 18.5 cm, L × W × H), which was filled with sawdust (5 cm) and had 20 equally spread marbles placed on top (5 x 4 rows). After mice had spent 30 minutes in the cage, the number of buried marbles was counted by two researchers and averaged. A buried marble was defined as 2/3 of the

marble not being visible anymore. Sawdust was renewed, and marbles cleaned with 70% ethanol in-between animals.

3-Chamber social interaction test

The three-chamber sociability test was used to assess social preference and recognition and was conducted as previously described (27). The testing apparatus was a three-chambered, rectangular box. The dividing walls between each chamber (20 × 40 × 22 cm, L × W × H) had small circular openings (5 cm diameter), allowing for access to all chambers. The two outer chambers contained wire cup-like cages (10 cm bottom diameter, 13 cm height), allowing for auditory, olfactory and visual, but not physical contact. The test consisted of 10-minute three phases: 1) Habituation, 2) Social preference, 3) Social recognition. In the first phase (Habituation), mice were allowed to explore the entire box with both wire cup-like cages left empty to allow for habituation to the novel environment. In the second phase (Social preference), one wire cup-like cage contained a novel, age-matched, conspecific, male mouse, whereas the other cage contained an object (rubber duckie). In the third phase (Social recognition), the mouse of the previous trial was left in the wire cup-like cage (Familiar mouse), while the object was replaced with a conspecific mouse (Novel mouse). The test mouse was held in the middle chamber while the conspecific mouse and object were placed in the cup wire-like cages. The location of the conspecific mice and object were systemically altered in-between test mice. The three-chamber test apparatus and wire cup-like cages were cleaned with 70% ethanol after each test mouse and left to dry for a few minutes. To reduce potential anxiogenic factors, all mice were habituated to the testing room 40 minutes before the test, the floor of the testing arena was covered with sawdust and testing was performed under dim light (60

lux). All experiments were videotaped using a ceiling camera and were scored blinded for the time interacted with the wire cup-like cages. The discrimination index was calculated as follows: (Time spent interacting with object or mouse/Total time spent interacting)*100%.

Elevated plus maze

The elevated plus maze test was used to assess anxiety-like behaviour and was conducted as previously described (6). The elevated plus maze apparatus was elevated 1 meter above the ground and consisted of a grey cross-shaped maze with two open arms and two closed arms (50 × 5 cm with 15 cm walls in the closed arms and 1 cm walls in the open arms). Mice were allowed to explore the maze for 5 min. Mice were habituated to the room 30 minutes prior to the test. Experiments were conducted in red light (5 lux). The elevated plus maze apparatus was cleaned with 70% ethanol in-between animals. Experiments were videotaped using a ceiling camera and videos were scored blinded for time spent in the open arms, which was defined as all paws in the open arm.

Open field test

Mice were assessed for locomotor activity and response to a novel environment in the open field test, which was conducted as previously described (6). Animals were placed in an open arena (40 × 32 × 24 cm, L × W × H) and were allowed to explore the arena for 10 minutes. Animals were habituated to the room 30 minutes prior to the test. Testing was performed under dim light (60 lux). The open field test box was cleaned with 70% ethanol in-between animals. Experiments were videotaped using a ceiling

camera and were analysed for time spend in the virtual centre zone (defined as 50% away from the edges) and total distance travelled using Ethovision version 13 software (Noldus).

Tail-suspension test

The tail-suspension test was used to assess depressive-like behaviour and was conducted as previously described (6). Mice were hung by their tail using adhesive tape (2 cm from the tip of the tail) to a 30 cm-elevated grid bar for 6 min. Experiments were videotaped using a numeric tripod-fixed camera and videos were scored blinded for the time mice spent immobile.

Saccharin preference test

Mice were assessed for reward-seeking behaviour using the saccharin preference test as previously conducted (28). Mice were first habituated to single housing and having two drinking water bottles for 3 days. Drinking water intake and food intake was measured during the habituation phase of the test. Hereafter, one drinking water bottle was replaced by one containing a saccharin solution (0.1% w/v) for 24 hours. Drinking water bottles were weighed every 12 hours during the testing phase to calculate saccharin preference. The side on which the regular drinking water bottle and the one containing saccharine solution was, were randomised and counterbalanced between groups. During the habituation phase, drinking water bottles were alternated every 24 hours, whereas bottles were alternated every 12 hours during the testing phase. Saccharin preference was calculated using the following formula: (Total Sucrose Intake/Total fluid intake)*100%.

Female urine sniffing test

Mice were assessed for hedonic and reward-seeking behaviour in the female urine sniffing test, which was performed as previously described (29). Prior to this experiment, vaginal smears from age-matched female C57Bl/6 mice (n=20; Envigo, UK) were taken and assessed for their estrous cycle. Urine from female mice in the esterus stage was collected and pooled. Male mice were habituated 45 min before the start of the test to the test room, with a cotton bulb attached to the lid of their housing cage. The test mice were subsequently introduced to a new cotton bulb containing 60 µl of sterile water. After a 45 min intertrial-interval, mice were introduced to a new cotton bulb containing 60 µl of urine from a female mouse in esterus for 3 min. The experiment was conducted in red light (5 lux). All tests were videotaped using a ceiling camera and interaction time with the cotton bulbs was scored blinded.

Stress-induced hyperthermia test

The stress-induced hyperthermia test was used to assess stress-responsiveness, which was conducted as previously described (6). Body temperature was determined at baseline (T1) and 15 minutes later (T2) by gently inserting a Vaseline-covered thermometer 2.0 cm into the rectum. The temperature was noted to the nearest 0.1 °C after it stabilised (~10 s). Mice were restrained by scruffing during this procedure which was the stressor. Animals were habituated to the testing room 1 hour prior to the test. The difference between T1 and T2 reflected the stress-induced hyperthermia.

Intestinal motility assay

Gastrointestinal motility was assessed as previously described (30). Briefly, mice were single-housed at 8.00 a.m. with *ad libitum* access to food and drinking water. Three hours later, 0.2 mL of non-absorbable 6% carmine red in 0.5% methylcellulose dissolved in sterile phosphate-buffered saline was administered by oral gavage, after which drinking water was removed. The latency for the excretion of the first red-coloured faecal pellet was subsequently timed as a measure of gastrointestinal motility.

Assessment of faecal water content and weight

Mice were single-housed for one hour during which faecal pellets were collected (± 9 per animal). Pellets were subsequently weighed, incubated at 50 °C for 24 hours and weighed again. The average weight per pellet and percentage of faecal water content were calculated.

Appetitive Y-maze

The appetitive Y-maze was used to assess long-term spatial learning and was performed as previously described (31). The test consisted of two phases; the initial learning phase, where the first associated between the location of the food reward and spatial reference cues were formed, and the reversal learning phase, where the location of the food reward was altered in reference to the spatial reference cues, in which the relearning of a context was measured.

The Y-maze apparatus was elevated 80 cm above the ground and consisted of three arms (50 x 9.5 cm, L x W, with a 0.5 cm-high rim) arranged at an angle of 120° of

each other (Figure S1A). The apparatus could be rotated during testing. A small plastic food well (a cap of a 15 mL tube) was placed at the distal end of each arm. Testing was performed under dim light (30 lux).

Prior to testing, mice were food restricted (3-4 gram food per day) and kept between 90-95% of their free-feeding bodyweight (Figure S1B). Two days later, animals were habituated in their home cage to the small plastic well containing 1 mL food reward (sweetened condensed milk diluted in water 1:1) per mouse before the onset of the active phase. Mice were subsequently habituated on the Y-maze apparatus in home cage groups until mice were freely running around and readily collecting the food reward (each arm contained 1 mL food reward), which took 2 days. Finally, mice were individually placed on the Y-maze until they were running and collected the food reward (each arm contained 0.1 mL food reward), which took 4 days.

During the first phase (Initial learning), mice were assigned a goal arm according to the position in the room, which was counter-balance between groups. The maze was rotated 120° every trial to prevent potential associations of the correct goal arm with the texture or smell of the arm. The starting position for each trial was determined by a pseudo-randomised computer sequence, which was different for each mouse but was the same across treatment groups. This sequence did not contain more than three consecutive starts in the same position to avoid temporary position preferences. Animals were tested in groups of eight, with four animals of two experimental groups (i.e. two home cages). Each mouse received ten trials per day with an inter-trial interval of approximately 10 minutes. The time of testing was counterbalanced between groups and rotated each day to reduce the effect of testing during a specific time of the day. Mice received eight consecutive days of initial learning, resulting in a total of 80 trials. During the second phase (reversal learning), the goal arm was

changed to a different arm, and the placement of the mice was changed accordingly. This phase lasted 5 days, resulting in a total of 50 trials.

For each trial, the food well on the goal arm was filled with 0.1 mL food reward (sweetened condensed milk diluted in water 1:1). The mouse was placed at the end of the start arm and was allowed to run freely on the maze. The entries into each arm were counted, as well as when the mouse went into the goal arm immediately, of which the latter was counted as a successful trial. The mouse was placed back into the home cage after it consumed the food reward. In the rare occasion that the mouse did not walk into the goal arm and collect the food reward within 90 seconds, then the mouse was gently guided towards the goal arm and given a chance to collect the food reward, after which it was also returned to the home cage. A trial where the mouse did not walk into any arm was excluded from the analysis, as this indicates that the mouse was anxious. An entry was counted when the tail of the animal passed the entry of the arm. Between mice, the food wells were not cleaned so that a slight odour of milk reward remained at all times, ensuring mice found the goal arm based on spatial cues, rather the olfactory cues.

Fear conditioning

Fear conditioning was used as a assess amygdala-dependent learning memory and was conducted as previously described (32). The test consisted of 3 days/phases; 1) Training, 2) Assessment of cued memory, 3) Assessment of contextual memory, each of which was carried on successive days with a 24-hour interval. In phase 1 (training), animals were recorded for 3 minutes (baseline), followed by 6 tone-conditioned stimuli (70 dB, 20 s), followed by a foot shock (0.6 mA, 2 s), with a 1-minute interval.

In phase 2 (Assessment of cued memory), mice were placed in a novel context (i.e. black-checker walls with a solid Plexiglas opaque floor, under which paper was placed containing a 400 µl vanilla solution (79.5% water/19.5% ethanol/1% vanilla-extract solution), and after an initial acclimation period of 2 minutes, mice received 40 presentations of the tone-conditioned stimuli, each lasting 30 seconds with a 5-second interval. In phase 3 (assessment of contextual memory), mice were placed in the context of day 1 and recorded for 5 minutes, without the presentation of any tone-conditioned stimuli. The fear conditioning apparatus was cleaned with 70% ethanol in-between animals.

Forced swim test

The forced swim test was used to assess depressive-like behaviour and was conducted as previously described (33). Mice were individually placed in a transparent glass cylinder (24 × 21 cm diameter), containing 15-cm-depth water (23-25 °C), for 6 minutes. Mice were gently dried after the test and water was renewed after each animal. Experiments were videotaped using a ceiling camera and videos were scored blinded for immobility time in the last 4 min of the test.

Tissue collection

Collection of faecal samples throughout the study was done by single housing mice until 2 pellets were dropped between 10.00 and 12.00 a.m. The order faecal pellet collection was counterbalanced between groups to minimise the effect of circadian rhythm. Pellets were snap-frozen on dry ice within 3 minutes after excretion and subsequently stored at -80 °C.

Animals were sacrificed by decapitation in a random fashion regarding test groups between 9.00 a.m. and 2.00 p.m. Trunk blood was collected in EDTA-containing tubes and 100 µl was put in a separate Eppendorf for flow cytometry. Both tubes were centrifuged for 10 min at 3,500 g at 4°C, after which plasma was collected and stored at −80 °C for later analysis. The remaining cell pellet of the Eppendorf containing 100 µl of blood was stored at 4 °C and subsequently used for flow cytometry. Mesenteric lymph nodes (MLNs) were dissected, cleaned from fat tissue and in stored in RPMI-1640 medium with L-glutamine and sodium bicarbonate (R8758, Sigma), supplemented with 10% FBS (F7524l, Sigma) and 1% Pen/strep (P4333, Sigma) at 4 °C for subsequent flow cytometry. The contents of the distal part of the ileum (2 cm), as well as faecal pellets, were collected, snap-frozen on dry ice and stored at −80 °C for later sequencing. The caecum was weighed, snap-frozen on dry ice and stored at −80 °C. The length of the colon was measured, and the proximal and distal 2 cm were collected and cut in half. One side was snap-frozen on dry ice and stored at −80 °C and the other treated with RNAlater (Sigma, R0901). This was done by incubating the tissues for 48 hours at 4°C, after which the RNAlater was removed and tissues were stored at -80 °C for later gene expression analysis. Whole brains were snap-frozen in ice-cold isopentane and stored at -80 °C.

Flow cytometry

Blood and MLNs collected when animals were sacrificed were processed on the same day for flow cytometry. Blood was resuspended in 10 mL home-made red blood cell lysis buffer (15.5 mM NH_4Cl, 1.2 mM $NaHCO_3$, 0.01 mM tetrasodium EDTA diluted in deionised water) for 3 minutes. Blood samples were subsequently centrifuged (1500 g, 5 minutes), split into 2 aliquots and resuspended in 45 µl staining buffer (autoMACS

Rinsing Solution (Miltenyi, 130-091-222) supplemented with MACS BSA stock solution (Miltenyi, 130-091-376)) for the staining procedure. MLNs were poured over a 70 µm strainer and disassembled using the plunger of a 1 mL syringe. The strainer was subsequently washed with 10 mL media (RPMI-1640 medium with L-glutamine and sodium bicarbonate, supplemented with 10% FBS and 1% Pen/strep), centrifuged and $1x10^6$ cells were resuspended in 45 µl staining buffer for the staining procedure. For the staining procedure, 5 µl of FcR blocking reagent (Miltenyi, 130-092-575) was added to each sample. Samples were subsequently incubated with a mix of antibodies (Blood aliquot 1; 5 µl CD11b-VioBright FITC (Miltenyi, 130-109-290), 5 µl LY6C-PE (Miltenyi, 130-102-391), 0.3 µl CX3CR1-PerCP-Cyanine5.5 (Biolegend, 149010) and 5 µl CCR2-APC (Miltenyi, 130-108-723); Blood aliquot 2 and MLNs; 1 µl CD4-FITC (ThermoFisher, 11-0042-82) and 1 µl CD25-PerCP-Cyanine5.5 (ThermoFisher, 45-0251-80)) and incubated for 30 minutes on ice. Blood aliquot 1 was subsequently fixed in 4% PFA for 30 minutes on ice, whilst Blood aliquot 2 and MLNs underwent intracellular staining using the eBioscience™ Foxp3 / Transcription Factor Staining Buffer Set (ThermoFisher, 00-5523-00), according to the manufacturers' instructions, using antibodies for intracellular staining (2 µl FoxP3-APC (ThermoFisher, 17-5773-82) and 5 µl Helios-PE (ThermoFisher, 12-9883-42)). Fixed samples were resuspended in staining buffer and analysed the subsequent day on the BD FACSCalibur flow cytometry machine. Data were analysed using FlowJo (version 10). The investigated cell populations were normalised to PBMC levels.

HPLC analysis

The concentrations of serotonin (5-HT) and 5-hydroxyindoleacetic acid (5HIAA) in ileal and colonic tissues were determined using high performance liquid

chromatography (HPLC), based on a methodology described previously (34). See supplemental material for a detailed description of these methods.

Statistical analysis on behavioural and physiological parameters in mice

All behavioural and physiological data were assessed for normality using the Shapiro-Wilk test and Levene's test for equality of variances. No gavage and milk gavage datasets were assessed for statistical significance using an unpaired Student's t-test to investigate the impact of milk gavage. The effect of kefir was determined by a two-way ANOVA, followed by Dunnett's post hoc test whenever data were normally distributed. If data were non-parametrically distributed, then a Kruskal-Wallis test, followed by a Mann-Whitney U test was used. Parametric data is depicted as bar graphs with points as individual datapoint and expressed as mean ± SEM. Non-parametric data is depicted as a box with whiskers plot. Statistical analysis was performed using SPSS software version 24 (IBM Corp). A p-value < 0.05 was deemed significant. Table S1 summarises all tests performed, in addition to their corresponding p-values.

DNA extractions and sequencing

For analysis of the kefir microbiome, DNA was extracted from the fermented milk using the PowerSoil DNA Isolation Kit, as described previously (35). For analysis of the murine gut microbiome, DNA was extracted from the total ileal contents, cecal contents and faecal pellets using the QIAamp PowerFecal DNA Kit. Whole-metagenome shotgun libraries were prepared using the Nextera XT kit in accordance with the Nextera XT DNA Library Preparation Guide from Illumina, with the

exception that tagmentation time was increased to 7 minutes. Kefir libraries were sequenced on the Illumina MiSeq sequencing platform with a 2 x 300 cycle v3 kit. Gut libraries were sequenced were sequenced on the Illumina NextSeq 500 with a NextSeq 500/550 High Output Reagent Kit v2 (300 cycles). All sequencing was performed at the Teagasc sequencing facility in accordance with standard Illumina sequencing protocols.

Bioinformatics

Murine reads were removed from the raw sequencing files using the NCBI Best Match Tagger (BMTagger) (ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/), and fastq files were converted to unaligned bam files using SAMtools (36). Duplicate reads were subsequently removed using Picard Tools (https://github.com/broadinstitute/picard). Next, low quality reads were removed using the trimBWAstyle.usingBam.pl script from the Bioinformatics Core at UC Davis Genome Center (https://github.com/genome/genome/blob/master/lib/perl/Genome/Site/TGI/Hmp/HmpSraProcess/trimBWAstyle.usingBam.pl). Specifically, MiSeq reads were filtered to 200 bp, while NextSeq 105 bp. All reads with a quality score less than Q30 were discarded. The resulting fastq files were then converted to fasta files using the fq2fa option from IDBA-UD (37).

Compositional analysis was performed using MetaPhlAn2 (38). Strain-level metagenomic analysis was performed using StrainPhlAn (26), which phylogenetically characterises strains by identifying single nucleotide polymorphisms in species-specific marker genes, and PanPhlAn (39), which

functionally characterises strains by aligning reads against a species-specific pangenome database. StrainPhlAn outputs were visualised using GraPhlAn (40). Custom PanPhlAn databases were constructed from complete genome assemblies which were annotated using Prokka (41). See Table S2 for the list of reference genomes used in this study. Functional analysis was performed with HUMAnN2 (42), using the --bypass-translated-search option, and PanPhlAn. HUMAnN2 measures the abundances of UniRef clusters (43) by aligning sequences against the ChocoPhlAn database. HUMAnN2 gene families were mapped to level-4 enzyme commission (EC) categories using HUMAnN2 utility mapping files.

Sequence data have been deposited in the European Nucleotide Archive (ENA).

Statistical analysis of shotgun metagenomic data

The R package vegan (44) was used for alpha diversity analysis and principal component analysis. The Wilcoxon rank-sum test was used to measure statistical differences in alpha diversity between groups, and p-values were adjusted using the Benjamini-Hochberg method. The adonis function from vegan was used for PERMANOVA (PERMutational ANalysis Of VAriance) analysis. The linear discriminant analysis (LDA) effect size (LEfSe) method (45) was used to investigate if any taxa or HUMAnN2 pathways were differentially abundant (i.e. LDA>3.0) between groups. Correlation analysis was performed using HAllA (https://bitbucket.org/biobakery/halla/wiki/Home). Data was visualised using hclust2 (https://bitbucket.org/nsegata/hclust2), GraPhlAn, and the R package ggplot2 (46).

RESULTS

The fermented milk drink kefir is well-tolerated

Kefir administration did affect body weight, body composition, food intake and drinking water intake (Figure S2A-F). In addition, no differences were found in basal body temperature, as detected in the stress-induced hyperthermia test, as well as the locomotor activity assessed in the open field test (Figure S2G, H). Overall, this indicates that the fermented milk drink kefir was well-tolerated by mice.

Kefir did not affect gastrointestinal motility

Assessment of gastrointestinal motility by carmine red administration showed that kefir did not induce any changes in gastrointestinal propulsion (Figure S3A). In line with these findings was the absence of differences in faecal pellet weight and faecal water content (Figure S3B, C). Finally, no differences in caecum weight and colon length were detected at the end of the study (Figure S3D, E).

Kefir modulates anxiety- and depressive-like, as well as reward-seeking behaviour

In the marble burying test, we found that administration of UK4 decreased the number of marbles buried ($F(2,35) = 5.464$, $p = 0.009$) (Figure 2A). Even though no

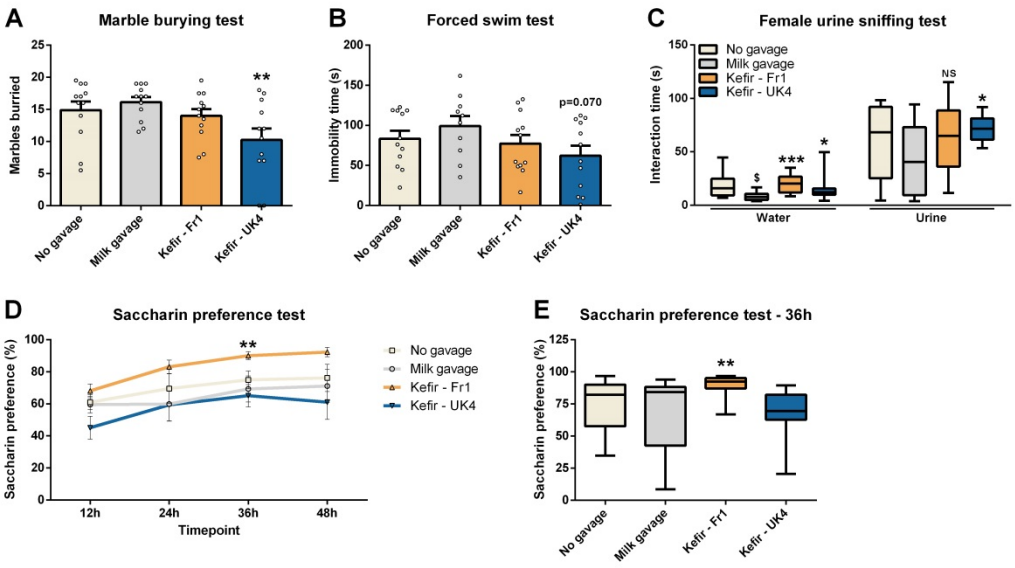


Figure 2: Kefir differentially affects repetitive/anxiety-like, depressive-like and reward-seeking behaviours. Repetitive/anxiety-like behaviour was assessed using the marble burying test (A). Depressive-like behaviour was determined using the forced swim test (B). Anhedonia and reward-seeking behaviours were investigated using the female urine sniffing test (C) and saccharin preference test (D, E). Significant differences are depicted as: *p < 0.05, **p < 0.01, ***p < 0.001 for Milk gavage compared to Kefir supplementation; and $p < 0.05 for No gavage compared to Milk gavage. All data are expressed as mean ± SEM (n = 11-12).

changes were observed in other tests assessing anxiety-like behaviour such as the elevated plus maze, open field test and stress-induced hyperthermia test (Figure S4A-C). Regarding depressive-like behaviour, UK4 induced a subtle trend towards decreased time spent immobile in the forced swim test ($F(2,33) = 2.327$, $p = 0.114$) (Figure 2B), even though this effect was not observed in the tail suspension test (Figure S4D). In the female urine sniffing test, mice receiving milk gavage spent less time interacting with the cotton bulb containing water compared to mice receiving no gavage ($\chi^2(1) = 6.367$, $p = 0.012$), which was ameliorated by both Fr1 and UK4 ($\chi^2(2) = 13.238$, $p < 0.001$) (Figure 2C). In addition, mice receiving UK4 spent more time interacting with the cotton bulb containing the urine from a female mouse in esterus, as a measure of reward-seeking behaviour ($\chi^2(2) = 6.280$, $p = 0.043$) (Figure 2C). Finally, Fr1 administration increased saccharin preference in the saccharin preference test, also often used as a measure of reward-seeking behaviour ($\chi^2(2) = 12.826$, $p = 0.002$) (Figure 2D, E).

Kefir does not affect sociability

All groups exhibited normal social preference and recognition in the 3-chamber social interaction test, indicating that kefir did not affect sociability (Figure S5A, B).

Kefir – UK4 modulates contextual learning and memory

No differences were observed in the fear conditioning test in phase 1 – acquisition, as determined by the time mice spent frozen during the presentation of the cue, as well is in-between the cues (Figure 3A, B). In addition, no differences were seen during phase 2, when cued-dependent fear memory was assessed (Figure 3C).

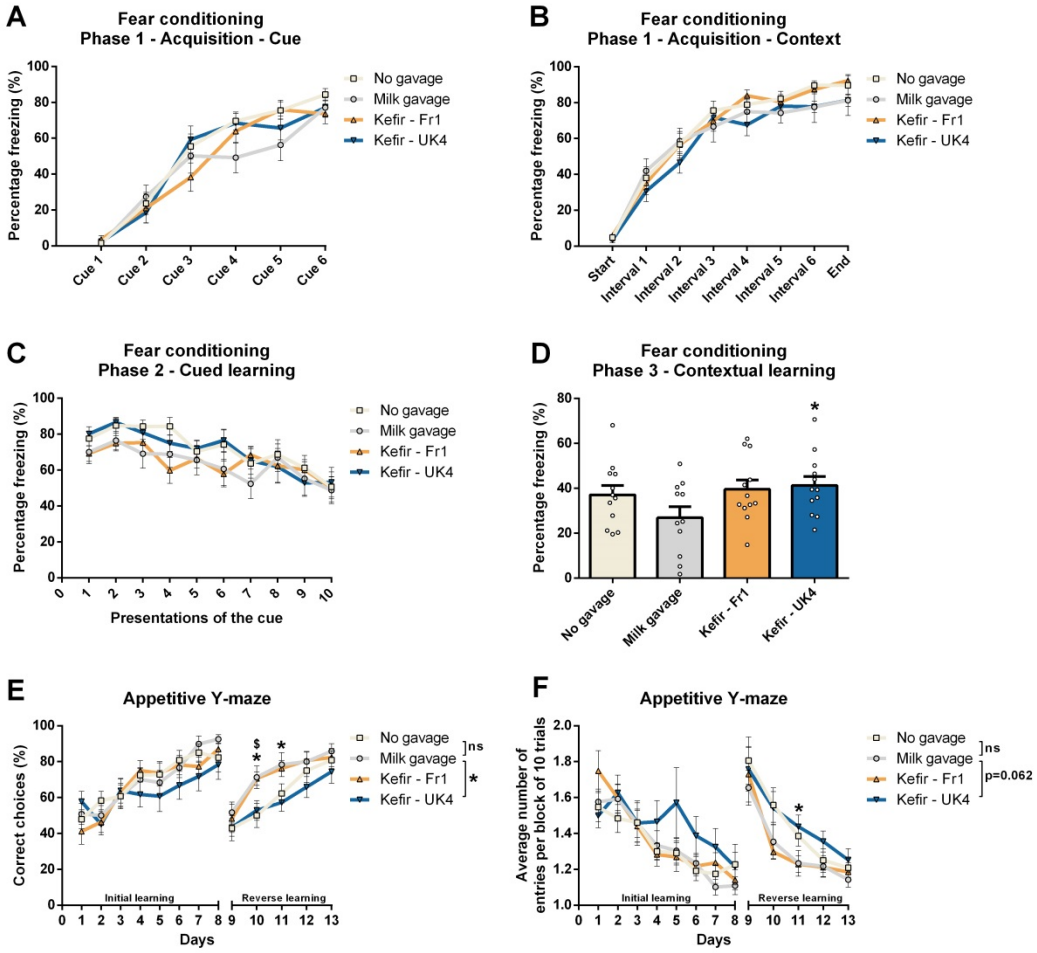


Figure 3: UK4 enhances fear-dependent contextual memory yet decreases long-term spatial learning. Fear-dependent memory and learning were assessed using fear conditioning. At phase 1 – Acquisition, mice were presented with a tone, followed by a foot shock. Cue-associative learning was assessed by measuring freezing behaviour during the presentation of the tone (A), whereas context-associative learning was determined in-between tones (B). At phase 2 – Cued memory, mice received 40 presentations of the same cue (the first 10 are shown), without foot shock, in a different context, in which fear-dependent cued memory was assessed (C). At phase 3 – Contextual memory, mice were exposed to the same context as day one for 5 minutes and contextual memory was assessed (D). Long-term spatial learning was assessed in the appetitive Y-maze, as determined by the percentage of times the mice made the correct choice as the first choice for reaching the goal (food reward) (E), as well as the number of average entries it took the mice to reach the goal (F). Significant differences are depicted as: *$p < 0.05$ for Milk gavage compared to Kefir supplementation; and $^{\$}p < 0.05$ No gavage compared to Milk gavage. All data are expressed as mean ± SEM (n = 10-12).

However, mice receiving UK4 showed a trend towards increased freezing behaviour in phase 3 – contextual memory ($F(2,34) = 3.181$, $p = 0.055$) (Figure 3D). Conversely, mice receiving UK4 made more errors in the reverse learning phase of the appetitive Y-maze as seen by the percentage correct choices ($F(2,33) = 3.870$, $p = 0.031$) (Figure 3E), and the amount of entries mice needed to reach the food reward ($F(2,33) = 3.387$, $p = 0.046$) (Figure 3F). It is interesting to note however, that a similar difference was found on day 10 in the percentage correct choices made between the "No gavage" and "Milk gavage" group ($t(22) = -2.303$, $p = 0.031$), where the mice receiving milk gavage performed superior (Figure 3F).

Kefir – Fr1 selectively increases colonic serotonergic activity

We found that mice receiving milk gavage showed decreased ileal serotonin (5-HT) levels compared to mice receiving no gavage ($t(21) = 2.650$, $p = 0.015$) (Figure 4B). This resulted in an increased 5HIAA/5-HT ratio ($t(22) = 2.650$, $p < 0.001$) (Figure 4C), indicating an increased serotonin turnover and serotonergic activity. The exact opposite was seen in the colon, where the milk gavage induced a trend towards increased 5-HT levels ($t(22) = -1.937$, $p = 0.066$) (Figure 4E), whilst decreasing the 5HIAA/5-HT ratio ($t(22) = 2.907$, $p = 0.008$) (Figure 4F). Interestingly, this phenotype in the colon, but not in the ileum, was ameliorated by Fr1 (5HIAA/5-HT ratio: $F(2,35) = 9.026$, $p < 0.001$) (Figure 4E, F).

Both kefirs differentially impact the peripheral immune system

UK4 increased the prevalence of T regulatory cells (Treg; CD4+, CD25+, FoxP3+) ($F(2,34) = 8.709$, $p < 0.001$) (Figure 5A), a well-known anti-inflammatory T helper cell subset known to be induced by gut microbial metabolites (47). Interestingly,

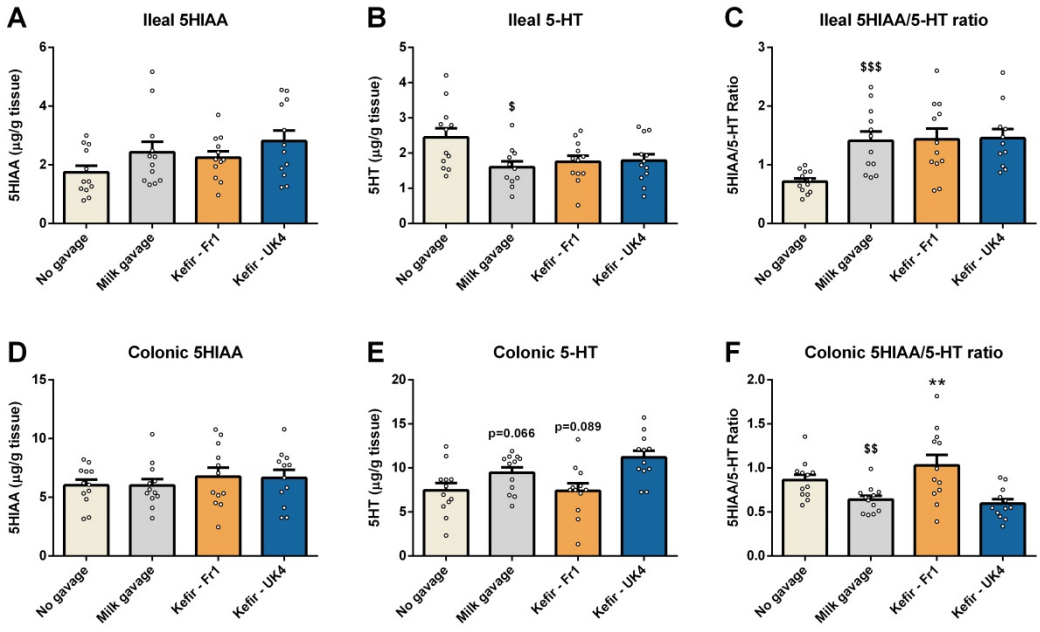


Figure 4: Fr1 modulates serotonergic signalling in the colon, but not ileum. Ileal (A-C) and colonic (D-F) tissues were quantified for 5HIAA and serotonin (5-HT) levels using HPLC. The 5HIAA/5-HT ratio was subsequently calculated. Significant differences are depicted as: **$p < 0.01$ for Milk gavage compared to Kefir supplementation; and $^{\$}p < 0.05$, $^{\$\$}p < 0.01$, $^{\$\$\$}p < 0.001$ for No gavage compared to Milk gavage. All data are expressed as mean ± SEM (n = 11-12).

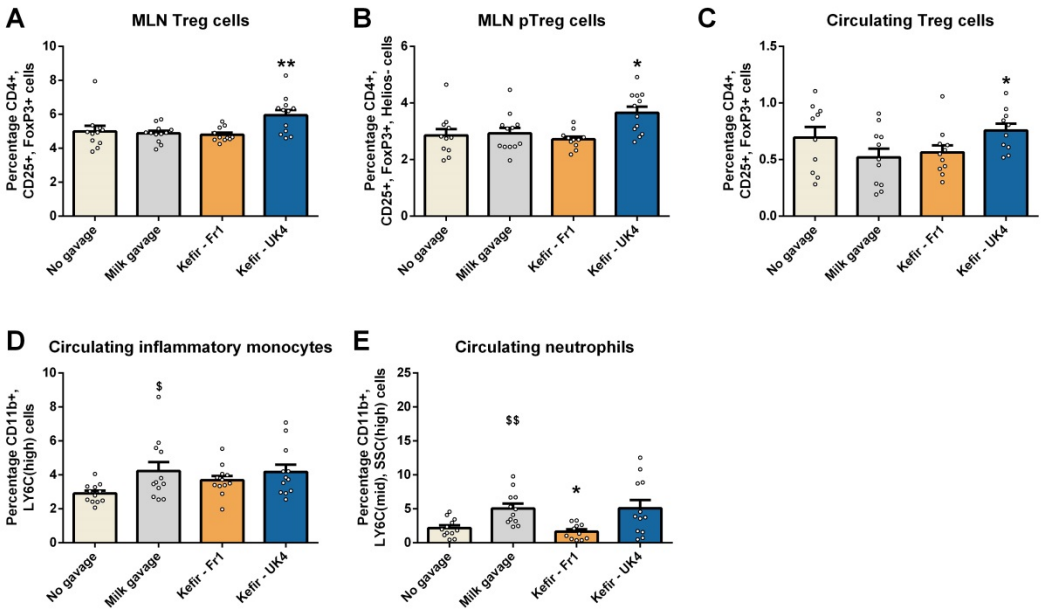


Figure 5: UK4 increases Treg cells levels, while Fr1 decreases neutrophil levels. Using flow cytometry, T regulatory cells (CD4+, CD25+, FoxP3+) were assessed in mesenteric lymph nodes (MLNs) and blood (A, C). Cells were subsequently assessed for Helios expression (B), as a measure of their origin (i.e. periphery (pTreg) or thymus). In addition, inflammatory monocytes (CD11b+, LY6C (high)) (D) and neutrophils (CD11b+, LY6C(mid), SSC(high)) (E) were assessed in the blood. Significant differences are depicted as: *$p < 0.05$, **$p < 0.01$ for Milk gavage compared to Kefir supplementation; and $^{\$}p < 0.05$, $^{\$\$}p < 0.01$ for No gavage compared to Milk gavage. All data are expressed as mean ± SEM (n = 11-12).

these cells did not express the Helios transcription factor ($F(2,34) = 7.548$, $p = 0.002$) (Figure 5B), indicating that they were induced in the periphery (pTreg) rather than in the thymus (48), potentially indicating that gut microbial-derived metabolites could have driven this increase in Treg cells. This UK4-induced increase Treg cells was also observed in the peripheral circulation ($F(2,31) = 3.420$, $p = 0.046$) (Figure 5C), indicating that these effects reached non-gastrointestinal tissues.

Interestingly, we observed an increase in circulating inflammatory monocytes (CD11b+, LY6C(high)) in mice receiving milk gavage, compared to mice receiving no gavage ($t(22) = -2.437$, $p = 0.023$) (Figure 5D). In line with this finding, was an increase in neutrophil levels (CD11b+, LY6C(mid), SSC(high)) induced by milk gavage ($t(22) = -3.583$, $p = 0.002$) (Figure 5E), indicating an activation of the innate immune system. The neutrophil levels however, were ameliorate by Fr1 administration ($F(2,34) = 5.412$, $p = 0.009$) (Figure 5E).

Kefir microbiota were largely stable over time

Shotgun metagenomics was used to determine the species-level composition of two kefirs, Fr1 and UK4, at six time-points over the twelve week duration of the experiment. Overall, the populations were temporally stable, and it was observed that both kefirs were dominated by *Lactococcus lactis,* and also consistently contained *Lactobacillus kefiranofaciens* (Figure S6). Several other species were identified but they were not consistently detected at each time-point. It was interesting to note that *Bifidobacterium breve* was detected at three time-points at >1% relative abundance in both kefirs. Additionally, *Pseudomonas* species, which likely originated in the

starting milk, were detected at the same two time-points in both kefirs at >10% relative abundance.

Kefirs exerted similar effects on gut microbiota composition, at both the species- and strain-levels

MetaPhlAn2 was employed to characterise the species-level microbial composition of the ileum, cecum and faeces. It was observed that the ileum was dominated by *Bifidobacterium pseudolongum* (73% in Fr1, 68% in UK4, and 56% in Milk), the cecum was dominated by *Mucispirillum schaedleri* (47% in Fr1, 40% in UK4, and 48% in Milk), while faeces were also dominated by *B. pseudolongum* (40% in Fr1, 35% UK4, and 29% in Milk) (Figure S7A). Additionally, *Lactobacillus* species, such as *Lactobacillus murinis* or *Lactobacillus reuteri*, were subdominant in each region. Expectedly, alpha diversity progressively increased from the ileum to the faeces (Figure S7B).

Pairwise comparisons were performed to identify differences between Fr1 versus Milk-fed mice, and UK4 versus Milk-fed mice. The Shannon diversity index was not significantly different between Fr1-fed mice versus Milk fed-mice in the ileum ($p=0.11$), cecum ($p=0.19$), or faeces ($p=0.16$) (Figure 6A). Similarly, PERMANOVA analysis indicated that there were no significant differences in beta diversity between Fr1-fed mice versus Milk-fed mice in the ileum ($p=0.088$, $R^2=0.111$), cecum ($p=0.087$, $R^2=0.087$), or faeces ($p=0.077$, $R^2=0.114$) (Figure 6B). However, LEfSe identified several differentially abundant species between Fr1-fed mice versus Milk-fed mice (Figure S8). In the ileum, *Bifidobacterium pseudolongum* (LDA=4.93) was significantly higher in Fr1-fed mice. In the cecum,

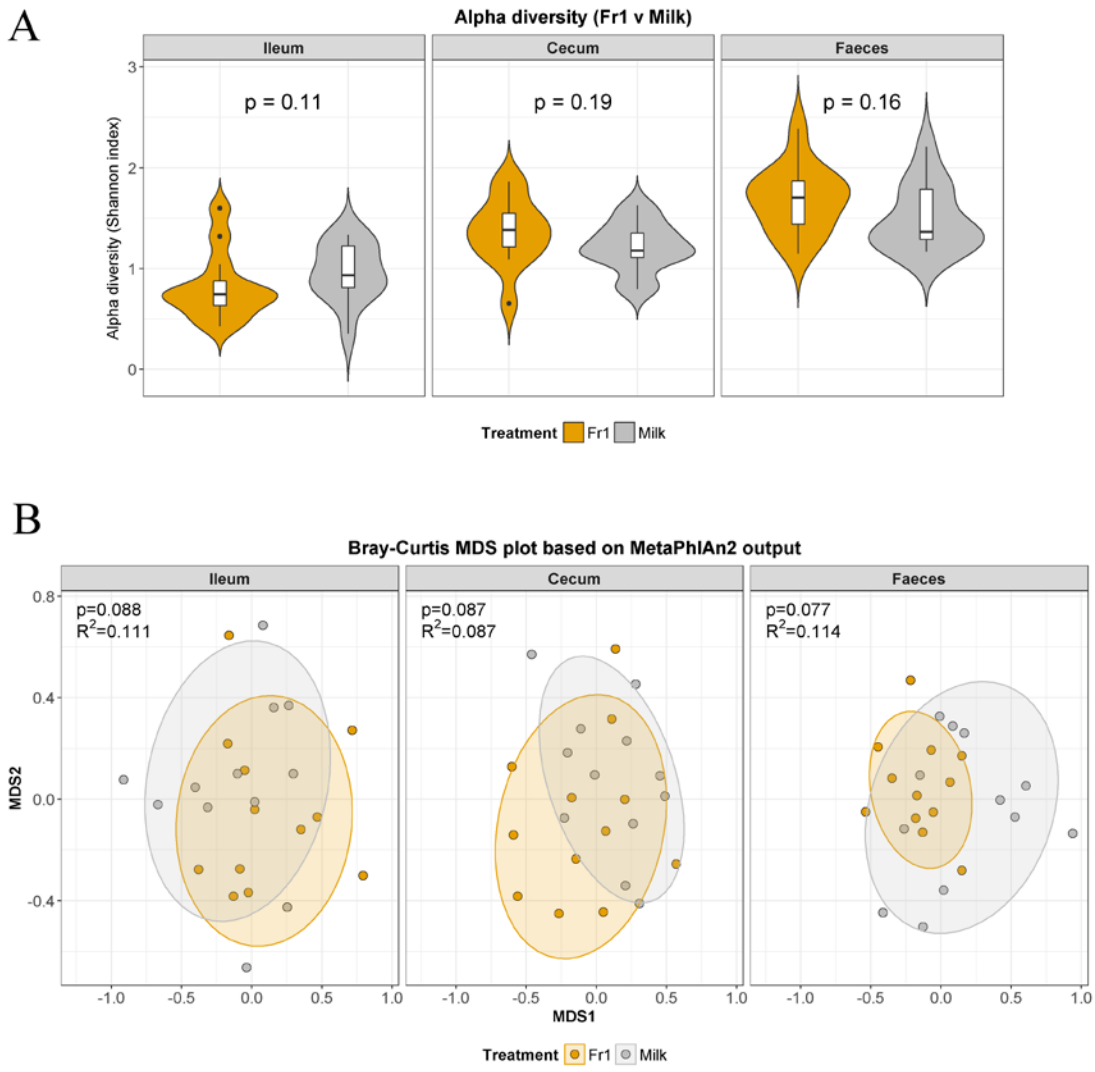


Figure 6: (A) Violin plots showing the alpha diversity (as measured using the Shannon index) of Fr1 versus Milk-fed mice. (B) MDS plots showing the dissimilarity in the microbial composition between Fr1 versus Milk-fed mice.

Parabacteroides goldsteinii (LDA=3.99) and *Lactobacillus reuteri* (LDA=4.36) were significantly higher in Fr1-fed mice, whereas Lachnospiraceae bacterium 3_1_46FAA (LDA=4.25) was significantly higher in Milk-fed mice. In the faeces, *Bacteroides intestinalis* (LDA=3.49), *Anaerotruncus* unclassified (LDA=3.75), *Eubacterium plexicaudatum* (LDA=3.77), and *Parabacteroides goldsteinii* (LDA=4.02) were significantly higher in Fr1-fed mice, whereas *Bacillus amyloliquefaciens* (LDA=3.04) and *Propionibacterium acnes* (LDA=3.25) were significantly higher in Milk-fed mice.

The Shannon diversity index was significantly higher in the cecum (p=0.017) in UK4 versus Milk-fed mice, but there were no significant differences in the ileum (p=0.44) or faeces (p=0.24) (Figure 7A). PERMANOVA analysis indicated that there were no significant differences in beta diversity between UK4 versus Milk-fed mice in the ileum (p=0.058, R^2=0.092), cecum (p=0.1, R^2=0.092), or faeces (p=0.073, R^2=0.09) (Figure 7B). LEfSe identified several differentially abundant species between UK4 versus Milk-fed mice (Figure S9). In the ileum, *Candidatus* Arthromitus unclassified (LDA=4.45) was higher in Milk-fed mice. In the cecum, *Alistipes* unclassified (LDA=4.08), *L. reuteri* (LDA=4.02), *Eubacterium plexicaudatum* (LDA=4.22) and *B. pseudolongum* (LDA=4.7) were higher in UK4-fed mice, whereas Lachnospiraceae bacterium 3_1_46FAA (LDA=4.28) was higher in Milk-fed mice. In the faeces, *E. plexicaudatum* (LDA=3.67) and *L. reuteri* (LDA=4.07) were higher in UK4-fed mice, whereas *B. amyloliquefaciens* (LDA=3.58) and *P. acnes* (LDA=4.04) were higher in Milk-fed mice.

Subsequently, PanPhlAn was used alongside StrainPhlAn to characterise differentially abundant species to the strain-level. Both tools indicated that the same *B. pseudolongum* strain, which was closely related to *B. pseudolongum* UMB-MBP-

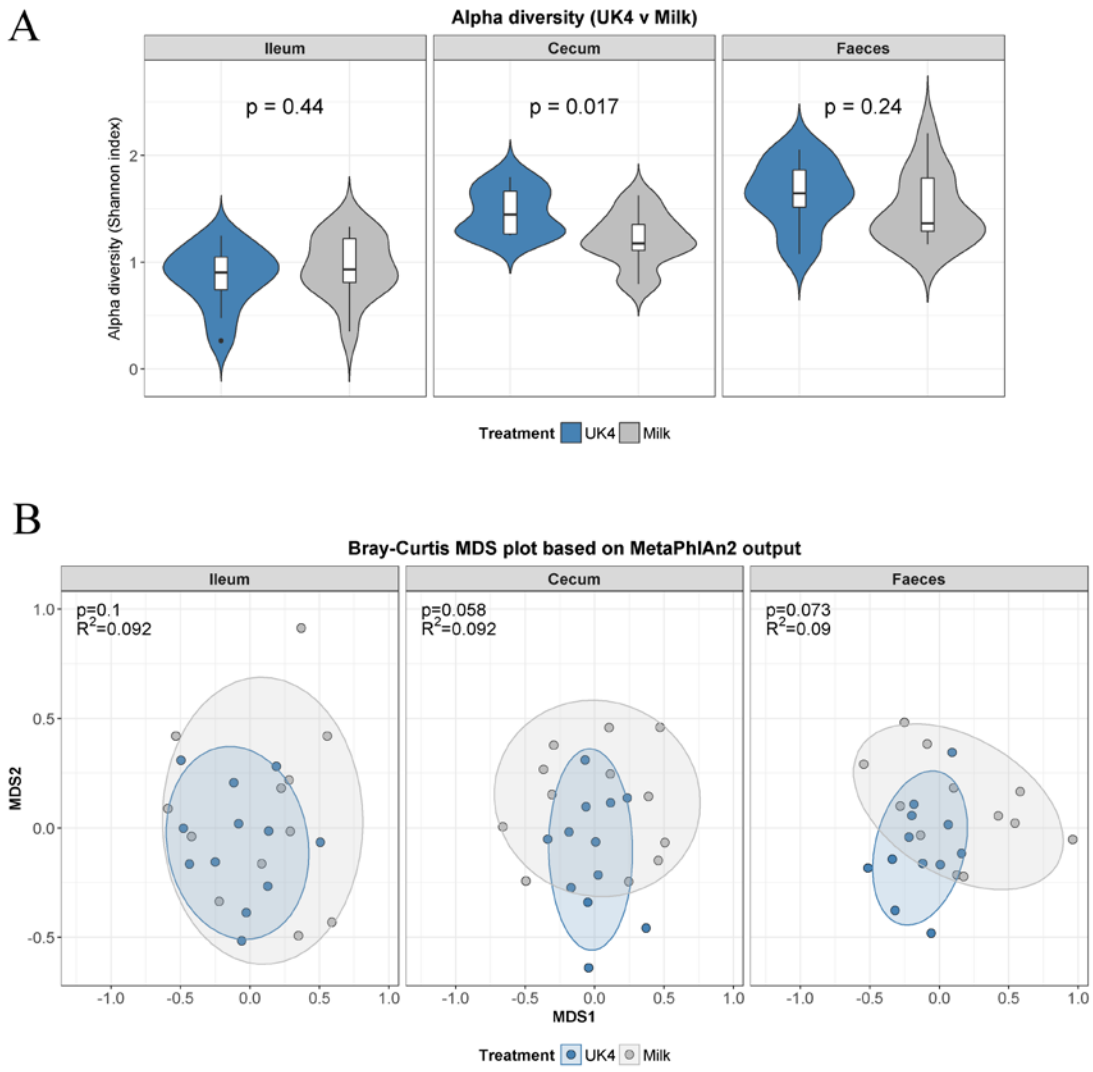


Figure 7: (A) Violin plots showing the alpha diversity (as measured using the Shannon index) of UK4 versus Milk-fed mice. (B) MDS plots showing the dissimilarity in the microbial composition between UK4 versus Milk-fed mice.

01, was present in each treatment group (Figure 8). Similarly, PanPhlAn indicated that the same *L. reuteri* strain, which was closely related to *L. reuteri* TD1, was also present in each treatment group (Figure 8). StrainPhlAn only detected a *L. reuteri* strain in one Fr1-fed sample, but again it indicated that this strain was closely related to *L. reuteri* TD1 (Figure 8). No other differentially abundant species could be characterised to the strain-level. Finally, neither PanPhlAn nor StrainPhlAn identified any of the strains detected in kefir in the gut microbiota.

Species relative abundances significantly correlate with immuno-physiological parameters in the murine gut

The tool HAllA revealed that no species were significantly associated with behavioural measurements (Figure S10), but several species were significantly associated with immuno-physiological parameters (Figure 9). In the ileum, *B. pseudolongum* was negatively associated with neutrophil levels (R=-0.52, q=0.47); Candidatus Arthromitus unclassified was negatively associated with Treg cell levels (R=-0.49, q=0.98); *Lactobacillus johnsonii* was positively associated with the ileal 5HIAA-5HT ratio (R=0.54, q=0.047); and *L. murinis* was positively associated with neutrophil levels (R=0.50, q=0.053). In the cecum, *L. johnsonii* was again positively associated with the ileal 5HIAA-5HT ratio (R=0.56, q=0.065). In the stool, there were no species were significantly associated with any immuno-physiological parameters.

Kefirs caused significant shifts in the functional potential of the gut microbiome

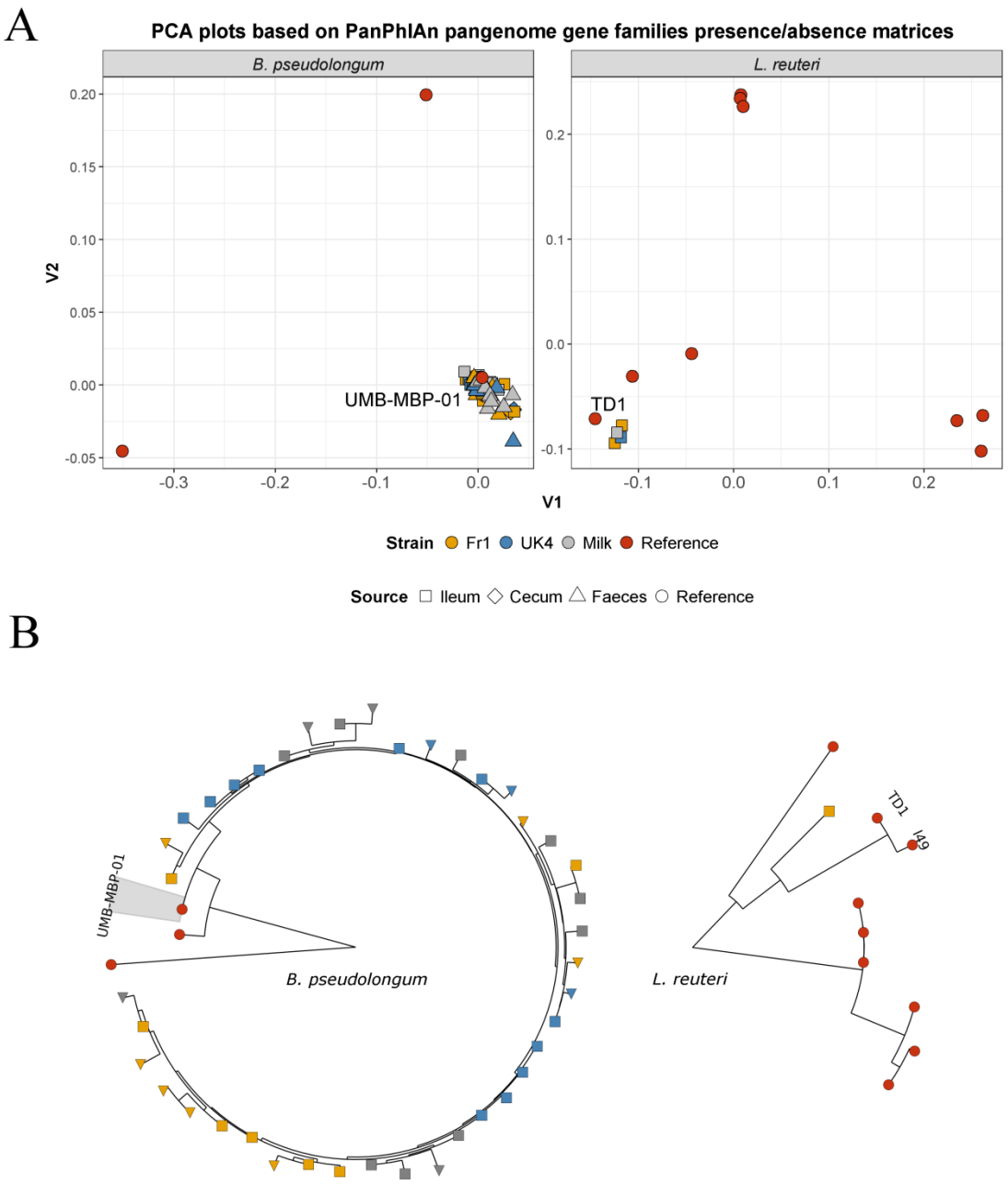


Figure 8: Strain-level analysis of *Bifidobacterium pseudolongum* and *Lactobacillus reuteri* detected in the mouse gut. (A) PCA plot based on gene families presence/absence matrices from PanPhlAn. The reference strains which shared the most gene families with that detected in the murine gastrointestinal tract are labelled. (B) Phylogenetic trees generated from StrainPhlAn outputs. Note that colours represent the group to which strains belong and shapes represent the source of the strains.

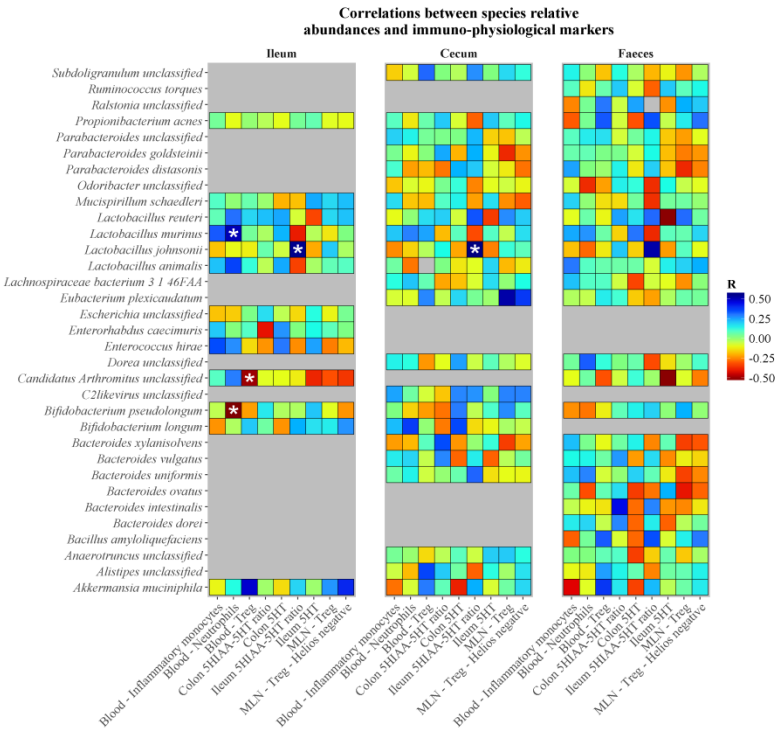


Figure 9: Correlations between species relative abundances and immuno-physiological parameters in the murine gut. The heatmap shows the Spearman rank correlation coefficient for each combination of variables. Significant associations, as determined by HAllA, are highlighted with asterisks.

HUMAnN2 was used to characterise the functional potential of the microbiome in the ileum, cecum and faeces. PERMANOVA analysis revealed that there was a significant functional separation in the microbiome between Fr1 versus Milk-fed mice in the cecum ($p=0.019$, $R^2=0.079$), but not in the ileum ($p=0.052$, $R^2=0.099$) or the faeces ($p=0.108$, $R^2=0.068$) (Figure 10A). Additionally, there was also a significant separation in the microbiome between UK4 versus Milk-fed animals in the cecum ($p=0.018$, $R^2=0.092$) and faeces ($p=0.01$, $R^2=0.09$), but not in the ileum ($p=0.212$, $R^2=0.092$) (Figure 10B). LEfSe identified 31 differentially abundant features (LDA>3.0) between Fr1 versus Milk-fed mice, while it also identified 23 differentially abundant features (LDA>3.0) between UK4 versus Milk-fed mice (Table S3). Notably, there were significant differences in several EC categories that may be involved in producing precursors to neurotransmitters. Specifically, genes encoding glutamine--fructose-6-phosphate transaminase (isomerising) (EC 2.6.1.16) (LDA=3.03), which produces glutamate, in addition to glutamate--ammonia ligase (EC 6.3.1.2) (LDA=3.34), which produces glutamine, were higher in the ileum in Fr1-fed mice compared to Milk-fed mice (Figure 10C). Genes encoding glutamine--fructose-6-phosphate transaminase (isomerising) were also higher in the ileum in UK4-fed mice compared to Milk-fed mice (LDA=3.31) (Figure 10C). Furthermore, the following EC categories were differentially abundant in the cecum in UK4-fed mice compared to Milk-fed mice: glutamate--ammonia ligase was higher in UK4-fed mice (LDA=3.1), whereas tryptophan synthase (EC 4.2.1.20), which produces tryptophan, was higher in Milk-fed mice (LDA=3.47) (Figure 10C). Finally, genes encoding glutamate--cysteine ligase (EC 6.3.2.2) were higher in the faeces in UK4-fed mice compared to Milk-fed mice (LDA=3.03) (Figure 10C).

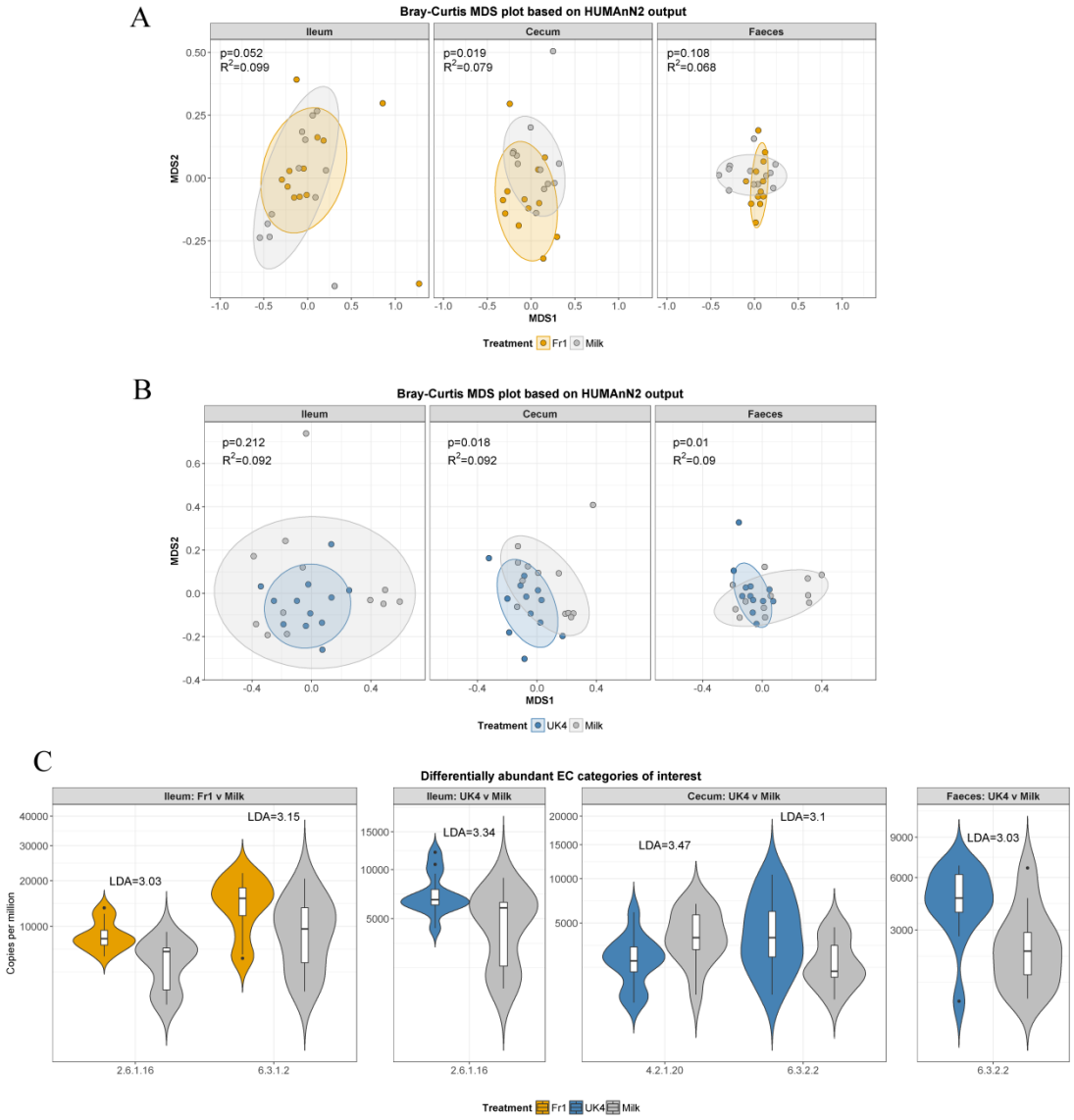


Figure 10: Functional analysis of the gut microbiome in mice fed kefir or unfermented milk. The MDS plots show the functional dissimilarity in the gut microbiome between (A) Fr1 versus Milk-fed mice and (B) UK4 versus Milk-fed mice. The violin plots (C) show differentially abundant EC level 4 categories of interest.

Potential GABA- and tryptophan-producing strains were increased following kefir ingestion

PanPhlAn gene-family matrices were examined to investigate if neurotransmitter-associated genes were present in either *B. pseudolongum* or *L. reuteri*, which were both significantly increased in the kefir groups. It was observed that the detected *B. pseudolongum* strain had genes encoding glutamate synthase, which may be involved in glutamate production, in addition to glutamine--fructose-6-phosphate transaminase (isomerising). Furthermore, it also encoded a putative glutamate/gamma-aminobutyrate antiporter, which may be involved in exporting glutamate or gamma-Aminobutyric acid (GABA) from the cell. However, no genes encoding glutamate decarboxylase, which produces GABA by the decarboxylation of glutamate, were identified in the detected *B. pseudolongum* strain. Interestingly, although HUMAnN2 indicated that tryptophan synthase (EC 4.2.1.20) was decreased in the kefir groups, genes encoding the protein were present in this strain.

The detected *L. reuteri* strain was found to encode glutamine--fructose-6-phosphate transaminase (isomerising). Importantly, this strain encoded glutamate decarboxylase along with a putative glutamate/gamma-aminobutyrate antiporter. Overall, these results indicate that both strains can potentially produce glutamate and/or GABA, while the detected *B. pseudolongum* strain might also synthesise tryptophan.

DISCUSSION

In the present study we report that, compared to unfermented fermented milk, two traditional kefirs, Fr1 and UK4, modulate repetitive and anxiety-like behaviour, depressive-like behaviour, reward-seeking behaviour, and cognition in mice.

We show the kefir UK4 induces an antidepressant-like effect in the forced swim test and female urine sniffing test, as well as a reduction in repetitive and anxiety-like behaviour in the marble burying test. In addition, UK4 enhanced contextual fear-dependent learning in the fear conditioning test, whilst decreasing long-term spatial learning in the appetitive Y-maze. The differing findings in regards to cognition and learning could indicate that UK4 modulates the central nervous system in a brain region-dependent manner, as fear-dependent learning is largely dependent on the involvement of the amygdala in contrary to long-term spatial learning in the appetitive Y-maze. The other kefir, Fr1, increased reward-seeking behaviour in the saccharin preference test and the female urine sniffing test as indicated by interaction time with the water-containing cotton bulb, indicating that Fr1 might be able to modulate central reward-circuitry. Interestingly, no significant differences were found in other behavioural tests assessing anxiety- and depressive-like behaviour as the elevated plus maze, open field test, stress-induced hyperthermia test and tail-suspension test. The differing findings regarding anxiety- and depressive-like behaviour across multiple tests highlight the benefits of screening in a battery style.

Shotgun metagenomics was employed to characterise the effects of the two kefirs on the ileal, cecal and faecal microbiome of mice. Species-level compositional analysis with MetaPhlAn2 revealed that both kefirs produced generally similar effects. Indeed, *B. pseudolongum* was increased in the ileum of Fr1-fed mice and the cecum of UK4-fed mice, *L. reuteri* was increased in the cecum of both groups and in the faeces of UK4-fed mice, while *E. plexicaudatum* was increased in the cecum of

UK4-fed mice and in the faeces of both groups. Conversely, Lachnospiraceae bacterium 3_1_46FAA was decreased in the cecum of both groups, while both *B. amyloliquefaciens* and *P. acnes* were decreased in the faeces of both groups. However, some kefir-specific effects were observed. Specifically, *P. goldsteinii* was increased in the cecum and faeces of Fr1-fed mice, while *Alistipes* unclassified was increased in the cecum of UK4-fed mice. Additionally, *Candidatus* Arthromitus unclassified was decreased in the ileum of UK4-fed mice. Furthermore, alpha diversity was only increased in the cecum of UK4-fed mice.

Several of these differentially abundant species have potential implications for health. Specifically, Lachnospiraceae, which was decreased in both kefir groups, has been frequently linked to obesity (49), whereas *P. goldsteinii*, which was increased in the Fr1 group, has been negatively correlated with this condition (50). Interestingly, although we did not observe any differences in percentage body fat here, two recent studies reported that kefir reduced weight gain in high-fat diet fed mice (18, 51). Furthermore, *B. pseudolongum* has been shown to increase the anti-inflammatory cytokine IL-10 in mice (52), while, similarly, *L. reuteri* has been shown to decrease inflammation in humans by increasing Treg cells (53). Importantly, UK4 increased anti-inflammatory Treg cells, suggesting that UK4 modulated the adaptive immune system, while Fr1 decreased neutrophils levels, suggesting that Fr1 modulated the innate immune system. It is possible that the increases in *B. pseudolongum* and/or *L. reuteri* contributed to the observed immune responses. Indeed, we identified a significant negative correlation between the abundance of *B. pseudolongum* in the ileum and blood neutrophil levels. Although it has already been established that kefir can modulate the immune system (11), this finding is particularly relevant here since immunomodulation by the gut microbiota

has been implicated in gut-brain axis signalling (54). This is further enforced by the observed changes in behaviour, even though more research is warranted to conclude any concrete mechanism. In addition, it should be noted that the observed immunomodulatory effects may also have been caused by metabolites in kefir itself, which were not studied here.

Strain-level analysis with PanPhlAn and StrainPhlAn indicated that the same *B. pseudolongum* and *L. reuteri* strains were present in each treatment group. Our finding suggests that these strains were endogenous to the gut, but that kefir promoted their growth. The detected *B. pseudolongum* strain was most closely related to *B. pseudolongum* UMB-MBP-01 (55), a strain that has been linked to improved organ transplant outcome in C57BL/6J mice, while the detected *L. reuteri* strain was most closely related to *L. reuteri* TD1 (56), a strain that was isolated from type 1 diabetes-resistant rats, which further indicates potential immunomodulatory roles for these bacteria. It is notable, though, that we did not detect any kefir strains in the mice, which suggests that the ingested microbes did not colonise the gut.

Metabolic reconstruction with HUMAnN2 revealed that both kefirs significantly altered the functional potential of the cecum, while UK4 also altered it in the faeces. Overall, across each region, 31 pathways were differentially abundant between Fr1 versus Milk-fed mice, while 23 pathways were differentially abundant between UK4 versus Milk-fed mice. Intriguingly, several genes involved in neurotransmitter production were differentially abundant between the groups. Specifically, genes encoding glutamine--fructose-6-phosphate transaminase (isomerising), which produces glutamate, were higher in the ileum of both kefir groups, while genes encoding glutamate--ammonia ligase, which produces glutamine, a precursor to glutamate, were also higher in the ileum of Fr1-fed mice. Glutamate, which is an

important excitatory neurotransmitter in the brain (57), is itself a precursor to *gamma*-aminobutyric acid (GABA), which is the main inhibitory neurotransmitter in the central nervous system (58). Subsequent strain-level functional analysis with PanPhlAn showed that both of the detected *B. pseudolongum* and *L. reuteri* strains encoded glutamine--fructose-6-phosphate transaminase (isomerising), while *B. pseudolongum* additionally encoded glutamate synthase. Furthermore, the detected *L. reuteri* strain encoded glutamate decarboxylase, which produces GABA from glutamate, while both strains were found to encode a putative glutamate/GABA antiporter. Thus, kefir consumption apparently increased the capacity for the gut microbiome to synthesise glutamate and/or GABA. Deficiencies in the GABA system have been linked to anxiety and depression (59). Interestingly, Bravo *et al.* previously showed that a probiotic, *Lactobacillus rhamnosus* JB-1, regulated emotional behaviour in mice by altering GABA receptor expression in the animals (60). Therefore, it is remarkable that anxious or depressive-like behaviours were decreased in both kefir groups. Our results might suggest that kefir reduced these symptoms by increasing GABA production in the gut.

HUMAnN2 also indicated that genes encoding tryptophan synthase, which produces tryptophan, were decreased in the cecum of UK4-fed mice. However, PanPhlAn showed that the detected *B. pseudolongum* strain, which was significantly increased in both kefir groups, encoded tryptophan synthase. Thus, UK4 consumption apparently decreased the total capacity for the microbiome to synthesise tryptophan in the cecum, but both kefirs increased a tryptophan producer in the same region. Indeed, this seemingly counterintuitive observation emphasises the value of strain-level functional analysis of the gut. Tryptophan is a precursor to the neurotransmitter serotonin, which is central to mood regulation in addition to cognition (61). It is

unclear if microbial synthesis of tryptophan influences host tryptophan levels, but it is noteworthy that, in a previous study, treatment with *Bifidobacterium infantis* was found to increase tryptophan levels in rats (62). Here, we observed that Fr1 increased serotonergic activity in the colon, but not in the ileum. It is possible that the detected *B. pseudolongum* strain contributed to this increase by augmenting tryptophan levels in the mice. Alternatively, serotonergic activity in the colon might have simply been increased because tryptophan is typically higher in kefir than in unfermented milk (13, 63).

In conclusion, the present study provides evidence which indicates that the traditional fermented dairy beverage kefir may modulate the gut-brain axis in mice. Our work supports the recent broadening of the definition of psychobiotic to include fermented foods like the fermented milk drink kefir. We show that kefir modulates repetitive and anxiety-like behaviour, depressive-like behaviour, reward-seeking behaviour, and cognition, while simultaneously increasing the abundance of bacterial strains containing genes associated with the biosynthesis of glutamate, GABA and tryptophan. However, it is possible that metabolites within kefir itself contributed to the observed improvements in behaviour, and therefore future investigations must address the effects of kefir isolates on mood. Regardless, our work suggests that kefir may serve as a dietary intervention to improve mood, and it merits further studies to confirm these effects in humans.

References

1. **Mayer EA, Knight R, Mazmanian SK, Cryan JF, Tillisch K.** 2014. Gut Microbes and the Brain: Paradigm Shift in Neuroscience. The Journal of Neuroscience **34:**15490.
2. **Bruce-Keller AJ, Salbaum JM, Berthoud H-R.** 2018. Harnessing Gut Microbes for Mental Health: Getting From Here to There. Biological Psychiatry **83:**214-223.
3. **Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, Morelli L, Canani RB, Flint HJ, Salminen S, Calder PC, Sanders ME.** 2014. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nature Reviews Gastroenterology &Amp; Hepatology **11:**506.
4. **Sheth RU, Cabral V, Chen SP, Wang HH.** 2016. Manipulating bacterial communities by in situ microbiome engineering. Trends in Genetics **32:**189-200.
5. **Sarkar A, Lehto SM, Harty S, Dinan TG, Cryan JF, Burnet PWJ.** 2016. Psychobiotics and the Manipulation of Bacteria–Gut–Brain Signals. Trends in Neurosciences **39:**763-781.
6. **Burokas A, Arboleya S, Moloney RD, Peterson VL, Murphy K, Clarke G, Stanton C, Dinan TG, Cryan JF.** 2017. Targeting the Microbiota-Gut-Brain Axis: Prebiotics Have Anxiolytic and Antidepressant-like Effects and Reverse the Impact of Chronic Stress in Mice. Biol Psychiatry **82:**472-487.
7. **Hilimire MR, DeVylder JE, Forestell CA.** 2015. Fermented foods, neuroticism, and social anxiety: An interaction model. Psychiatry Research **228:**203-208.

8. **Miyake Y, Tanaka K, Okubo H, Sasaki S, Arakawa M.** 2014. Intake of dairy products and calcium and prevalence of depressive symptoms during pregnancy in Japan: a cross-sectional study. BJOG: An International Journal of Obstetrics & Gynaecology **122:**336-343.

9. **Tillisch K, Labus J, Kilpatrick L, Jiang Z, Stains J, Ebrat B, Guyonnet D, Legrain–Raspaud S, Trotin B, Naliboff B.** 2013. Consumption of fermented milk product with probiotic modulates brain activity. Gastroenterology **144:**1394-1401. e1394.

10. **Leite AMO, Miguel MAL, Peixoto RS, Ruas-Madiedo P, Paschoalin VMF, Mayo B, Delgado S.** 2015. Probiotic potential of selected lactic acid bacteria strains isolated from Brazilian kefir grains. Journal of Dairy Science **98:**3622-3632.

11. **Bourrie BCT, Willing BP, Cotter PD.** 2016. The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir. Frontiers in Microbiology **7**.

12. **de Oliveira Leite AM, Miguel MA, Peixoto RS, Rosado AS, Silva JT, Paschoalin VM.** 2013. Microbiological, technological and therapeutic properties of kefir: a natural probiotic beverage. Braz J Microbiol **44:**341-349.

13. **Rosa DD, Dias MMS, Grześkowiak ŁM, Reis SA, Conceição LL, Peluzio MdCG.** 2017. Milk kefir: nutritional, microbiological and health benefits. Nutr Res Rev **30:**82-96.

14. **Rodrigues KL, Caputo LRG, Carvalho JCT, Evangelista J, Schneedorf JM.** 2005. Antimicrobial and healing activity of kefir and kefiran extract. International Journal of Antimicrobial Agents **25:**404-408.

15. **Liu JR, Wang SY, Chen MJ, Yueh PY, Lin CW.** 2006. The anti-allergenic properties of milk kefir and soymilk kefir and their beneficial effects on the intestinal microflora. Journal of the Science of Food and Agriculture **86:**2527-2533.

16. **Lee M-Y, Ahn K-S, Kwon O-K, Kim M-J, Kim M-K, Lee I-Y, Oh S-R, Lee H-K.** 2007. Anti-inflammatory and anti-allergic effects of kefir in a mouse asthma model. Immunobiology **212:**647-654.

17. **Kim DH, Kim H, Jeong D, Kang IB, Chon JW, Kim HS, Song KY, Seo KH.** 2017. Kefir alleviates obesity and hepatic steatosis in high-fat diet-fed mice by modulation of gut microbiota and mycobiota: targeted and untargeted community analysis with correlation of biomarkers. J Nutr Biochem **44:**35-43.

18. **Bourrie BCT, Cotter PD, Willing BP.** 2018. Traditional kefir reduces weight gain and improves plasma and liver lipid profiles more successfully than a commercial equivalent in a mouse model of obesity. Journal of Functional Foods **46:**29-37.

19. **Sylvia KE, Demas GE.** 2018. A gut feeling: Microbiome-brain-immune interactions modulate social and affective behaviors. Hormones and Behavior **99:**41-49.

20. **Noori N, Bangash MY, Motaghinejad M, Hosseini P, Noudoost B.** 2014. Kefir protective effects against nicotine cessation-induced anxiety and cognition impairments in rats. Adv Biomed Res **3:**251.

21. **Quince C, Walker AW, Simpson JT, Loman NJ, Segata N.** 2017. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35:**833.

22. **Dovrolis N, Kolios G, Spyrou GM, Maroulakou I.** 2017. Computational profiling of the gut-brain axis: microflora dysbiosis insights to neurological disorders. Brief Bioinform doi:10.1093/bib/bbx154.

23. **Noecker C, McNally CP, Eng A, Borenstein E.** 2017. High-resolution characterization of the human microbiome. Translational Research **179:**7-23.

24. **Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C.** 2015. Sequencing and beyond: integrating molecular'omics' for microbial community profiling. Nature Reviews Microbiology **13:**360-372.

25. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. **13:**435-438.

26. **Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N.** 2017. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research **27:**626-638.

27. **Desbonnet L, Clarke G, Shanahan F, Dinan TG, Cryan JF.** 2014. Microbiota is essential for social development in the mouse. Mol Psychiatry **19:**146-148.

28. **O'Leary OF, Felice D, Galimberti S, Savignac HM, Bravo JA, Crowley T, El Yacoubi M, Vaugeois JM, Gassmann M, Bettler B, Dinan TG, Cryan JF.** 2014. GABAB(1) receptor subunit isoforms differentially regulate stress resilience. Proc Natl Acad Sci U S A **111:**15232-15237.

29. **Finger BC, Dinan TG, Cryan JF.** 2011. High-fat diet selectively protects against the effects of chronic social stress in the mouse. Neuroscience **192:**351-360.

30. **Golubeva AV, Joyce SA, Moloney G, Burokas A, Sherwin E, Arboleya S, Flynn I, Khochanskiy D, Moya-Perez A, Peterson V, Rea K, Murphy K, Makarova O, Buravkov S, Hyland NP, Stanton C, Clarke G, Gahan CGM, Dinan TG, Cryan JF.** 2017. Microbiota-related Changes in Bile Acid & Tryptophan Metabolism are Associated with Gastrointestinal Dysfunction in a Mouse Model of Autism. EBioMedicine **24:**166-178.

31. **Finger BC, Dinan TG, Cryan JF.** 2010. Leptin-deficient mice retain normal appetitive spatial learning yet exhibit marked increases in anxiety-related behaviours. Psychopharmacology (Berl) **210:**559-568.

32. **Izquierdo A, Wellman CL, Holmes A.** 2006. Brief uncontrollable stress causes dendritic retraction in infralimbic cortex and resistance to fear extinction in mice. J Neurosci **26:**5733-5738.

33. **Cryan JF, Mombereau C.** 2004. In search of a depressed mouse: utility of models for studying depression-related behavior in genetically modified mice. Mol Psychiatry **9:**326-357.

34. **Browne CA, Clarke G, Dinan TG, Cryan JF.** 2011. Differential stress-induced alterations in tryptophan hydroxylase activity and serotonin turnover in two inbred mouse strains. Neuropharmacology **60:**683-691.

35. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.

36. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

37. **Peng Y, Leung HC, Yiu S-M, Chin FY.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28:**1420-1428.

38. **Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N.** 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods **12:**902-903.

39. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods.

40. **Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N.** 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ **3:**e1029.

41. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068-2069.

42. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B.** 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Computational Biology **8:**e1002358.

43. **Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U.** 2014. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31:**926-932.

44. **Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M.** 2007. The vegan package. Community Ecology Package **10:**631-637.

45. **Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C.** 2011. Metagenomic biomarker discovery and explanation. Genome Biol **12:**R60.

46. **Wickham H.** 2016. ggplot2: elegant graphics for data analysis. Springer.

47. **Tanoue T, Atarashi K, Honda K.** 2016. Development and maintenance of intestinal regulatory T cells. Nat Rev Immunol **16:**295-309.

48. **Shevach EM, Thornton AM.** 2014. tTregs, pTregs, and iTregs: similarities and differences. Immunol Rev **259:**88-102.

49. **Meehan CJ, Beiko RG.** 2014. A Phylogenomic View of Ecological Specialization in the Lachnospiraceae, a Family of Digestive Tract-Associated Bacteria. Genome Biology and Evolution **6:**703-713.

50. **Chang C-J, Lin C-S, Lu C-C, Martel J, Ko Y-F, Ojcius DM, Tseng S-F, Wu T-R, Chen Y-YM, Young JD, Lai H-C.** 2015. Ganoderma lucidum reduces obesity in mice by modulating the composition of the gut microbiota. Nature Communications **6:**7489.

51. **Kim D-H, Kim H, Jeong D, Kang I-B, Chon J-W, Kim H-S, Song K-Y, Seo K-H.** 2017. Kefir alleviates obesity and hepatic steatosis in high-fat diet-fed mice by modulation of gut microbiota and mycobiota: targeted and untargeted community analysis with correlation of biomarkers. J Nutr Biochem **44:**35-43.

52. **Sasajima N, Ogasawara T, Takemura N, Fujiwara R, Watanabe J, Sonoyama K.** 2009. Role of intestinal Bifidobacterium pseudolongum in dietary fructo-oligosaccharide inhibition of 2,4-dinitrofluorobenzene-induced contact hypersensitivity in mice. British Journal of Nutrition **103:**539-548.

53. **Mu Q, Tavella VJ, Luo XM.** 2018. Role of Lactobacillus reuteri in Human Health and Diseases. Frontiers in Microbiology **9**.

54. **Stilling RM, Dinan TG, Cryan JF.** 2014. Microbial genes, brain & behaviour–epigenetic regulation of the gut–brain axis. Genes, Brain and Behavior **13:**69-86.

55. **Mongodin EF, Hittle LL, Nadendla S, Brinkman CC, Xiong Y, Bromberg JS.** 2017. Complete Genome Sequence of a Strain of Bifidobacterium pseudolongum Isolated from Mouse Feces and Associated with Improved Organ Transplant Outcome. Genome Announc **5**.

56. **Leonard MT, Valladares RB, Ardissone A, Gonzalez CF, Lorca GL, Triplett EW.** 2014. Complete Genome Sequences of Lactobacillus johnsonii Strain N6.2 and Lactobacillus reuteri Strain TD1. Genome Announc **2**.

57. **Meldrum BS.** 2000. Glutamate as a Neurotransmitter in the Brain: Review of Physiology and Pathology. The Journal of Nutrition **130:**1007S-1015S.

58. **Owens DF, Kriegstein AR.** 2002. Is there more to gaba than synaptic inhibition? Nature Reviews Neuroscience **3:**715.

59. **Möhler H.** 2012. The GABA system in anxiety and depression and its therapeutic potential. Neuropharmacology **62:**42-53.

60. **Bravo JA, Forsythe P, Chew MV, Escaravage E, Savignac HM, Dinan TG, Bienenstock J, Cryan JF.** 2011. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. Proceedings of the National Academy of Sciences **108:**16050.

61. **Canli T, Lesch K-P.** 2007. Long story short: the serotonin transporter in emotion regulation and social cognition. Nature Neuroscience **10:**1103.

62. **Desbonnet L, Garrett L, Clarke G, Kiely B, Cryan JF, Dinan TG.** 2010. Effects of the probiotic Bifidobacterium infantis in the maternal separation model of depression. Neuroscience **170:**1179-1188.

63. **Arslan S.** 2015. A review: chemical, microbiological and nutritional characteristics of kefir. CyTA - Journal of Food **13:**340-345.

Supplemental material

HPLC analysis

Mobile phase consisted of 0.1M citric acid, 0.1M sodium dihydrogen phosphate monohydrate, 0.01mM EDTA disodium salt (Alkem/Reagecon, Cork), 5.6mM octane-1-sulphonic acid (Sigma Aldrich), and 9% (v/v) methanol (Alkem/Reagecon). The pH of the mobile phase was adjusted to 2.8 using 4N sodium hydroxide (Alkem/Reagecon). Briefly, tissue samples were sonicated (Sonopuls HD 2070, Bandelin, Berlin Germany) in 500uL of cold mobile phase containing 4ng/40uL of N-methyl serotonin (Sigma Aldrich). Tissues were sonicated for 4 seconds and were kept chilled during sonication. Tissue homogenates were then centrifuged at 14000RPM for 20min at 4^{o}C. The supernatant was collected and transferred to a new collection tube; the pellet was discarded. The supernatant was then vortexed and 30 uL of supernatant was spiked into 270uL of mobile phase that did not contain N-methyl serotonin. 20uL of the 1:10 dilution (4^{o}C) was injected into the HPLC system (Shimadzu, Japan) which was comprised of a SCL 10-Avp system controller, LC-10AS pump, SIL-10A autoinjector, CTO-10A oven, LECD 6A electrochemical detector, and Class VP-5 software. The chromatographic conditions were flow rate of 0.9mL/min using a Kinetex 2.6u C18 100A x 4.6mm column, oven temperature of 30^{o}C, and detector settings of +0.8V. The total run time for each sample was 40min. External standards (serotonin creatinine sulfate and 5-hydroxyindole-3-acetic acid; Sigma Aldrich) were run in duplicate at a final concentration of 2ng/20uL. Monoamines in unknown samples were determined by their retention times compared to external standards. Peak heights of the analyte:internal standard ratio were used to quantitate monomamine concentrations in each sample. Monoamine concentration was presented as ug of monoamine per g of tissue.

Table S1. Summary of statistical analysis on behavioural and physiological parameters in mice. Note that NG represents "No gavage".

Test	Comparison	Measure	P-value
3-Chamber test - Social preference	Fr1 v Milk	Mouse	0.998
		Object	0.998
	NG v Milk	Mouse	0.745
		Object	0.729
	UK4 v Milk	Mouse	0.693
		Object	0.831
3-Chamber test - Social recognition	Fr1 v Milk	Familiar	0.685
		Novel	0.685
	NG v Milk	Familiar	0.745
		Novel	0.729
	UK4 v Milk	Familiar	0.693
		Novel	0.831
Appetitive Y-maze - %	Fr1 v Milk	Correct choices (%) D1	0.582
		Correct choices (%) D10	0.9996
		Correct choices (%) D11	0.9858
		Correct choices (%) D12	1
		Correct choices (%) D13	0.94
		Correct choices (%) D2	0.943
		Correct choices (%) D3	0.995
		Correct choices (%) D4	0.88
		Correct choices (%) D5	0.919
		Correct choices (%) D6	0.996
		Correct choices (%) D7	0.3065
		Correct choices (%) D8	0.876
		Correct choices (%) D9	0.969
	NG v Milk	Correct choices (%) D10	0.031
		Correct choices (%) D1	0.97
		Correct choices (%) D11	0.1119
		Correct choices (%) D12	0.872
		Correct choices (%) D13	0.844
		Correct choices (%) D2	0.588
		Correct choices (%) D3	1
		Correct choices (%) D4	0.981
		Correct choices (%) D5	0.943
		Correct choices (%) D6	0.932
		Correct choices (%) D7	0.8778
		Correct choices (%) D8	0.529
		Correct choices (%) D9	0.66
	UK4 v Milk	Correct choices (%) D11	0.031
		Correct choices (%) D1	0.766
		Correct choices (%) D10	0.0697
		Correct choices (%) D12	0.21

Test	Comparison	Measure	P-value
		Correct choices (%) D13	0.307
		Correct choices (%) D2	0.862
		Correct choices (%) D3	0.981
		Correct choices (%) D4	0.62
		Correct choices (%) D5	0.799
		Correct choices (%) D6	0.53
		Correct choices (%) D7	0.0779
		Correct choices (%) D8	0.273
		Correct choices (%) D9	0.694
Appetitive Y-maze - Average	Fr1 v Milk	Average number of entries: D1	0.411
		Average number of entries: D10	0.936
		Average number of entries: D11	0.9999
		Average number of entries: D12	1
		Average number of entries: D13	0.887
		Average number of entries: D2	1
		Average number of entries: D3	0.999
		Average number of entries: D4	0.949
		Average number of entries: D5	0.994
		Average number of entries: D6	0.998
		Average number of entries: D7	0.3829
		Average number of entries: D8	0.984
		Average number of entries: D9	0.953
	NG v Milk	Average number of entries: D1	0.99
		Average number of entries: D10	0.224
		Average number of entries: D11	0.1955
		Average number of entries: D12	0.959
		Average number of entries: D13	0.706
		Average number of entries: D2	0.689
		Average number of entries: D3	1
		Average number of entries: D4	0.984
		Average number of entries: D5	1
		Average number of entries: D6	0.956
		Average number of entries: D7	0.7822
		Average number of entries: D8	0.591
		Average number of entries: D9	0.739
	UK4 v Milk	Average number of entries: D11	0.046
		Average number of entries: D1	0.872
		Average number of entries: D10	0.286
		Average number of entries: D12	0.243
		Average number of entries: D13	0.325
		Average number of entries: D2	0.984
		Average number of entries: D3	1
		Average number of entries: D4	0.534

Test	Comparison	Measure	P-value
		Average number of entries: D5	0.26
		Average number of entries: D6	0.308
		Average number of entries: D7	0.0626
		Average number of entries: D8	0.675
		Average number of entries: D9	0.896
Body Temperature	Fr1 v Milk	Body temperature (°C)	0.25
	NG v Milk		0.994
	UK4 v Milk		0.374
Body weight	Fr1 v Milk	Body weight: D0	0.126
		Body weight: D100	0.58
		Body weight: D14	0.509
		Body weight: D20	0.987
		Body weight: D27	1
		Body weight: D3	0.0848
		Body weight: D34	1
		Body weight: D41	0.984
		Body weight: D48	0.994
		Body weight: D56	0.948
		Body weight: D64	0.66
		Body weight: D7	0.374
		Body weight: D92	0.229
	NG v Milk	Body weight: D0	0.983
		Body weight: D100	0.952
		Body weight: D14	0.117
		Body weight: D20	0.504
		Body weight: D27	0.609
		Body weight: D3	0.3357
		Body weight: D34	0.106
		Body weight: D41	0.314
		Body weight: D48	0.13
		Body weight: D56	0.157
		Body weight: D64	0.813
		Body weight: D7	0.102
		Body weight: D92	0.995
	UK4 v Milk	Body weight: D0	0.572
		Body weight: D100	0.474
		Body weight: D14	0.993
		Body weight: D20	0.935
		Body weight: D27	0.803
		Body weight: D3	0.7277
		Body weight: D34	0.844
		Body weight: D41	0.845
		Body weight: D48	0.98

Test	Comparison	Measure	P-value
		Body weight: D56	0.719
		Body weight: D64	0.791
		Body weight: D7	0.958
		Body weight: D92	0.249
Cecum weight	Fr1 v Milk	Cecum weight (%)	0.996
	NG v Milk		0.955
	UK4 v Milk		0.584
Circulating inflammatory monocytes	Fr1 v Milk	Percentage CD11b+, LY6C(high) cells	0.5932
	NG v Milk		0.023
	UK4 v Milk		0.9993
Circulating neutrophils	Fr1 v Milk	Percentage CD11b+, LY6C(mid), SSC(high) cells	0.009
	NG v Milk		0.002
	UK4 v Milk		0.9999
Circulating Treg cells	Fr1 v Milk	Percentage CD4+, CD25+, FoxP3+ cells	0.9496
	NG v Milk		0.2456
	UK4 v Milk		0.046
Colon 5HIAA	Fr1 v Milk	5HIAA (µg/g tissue)	0.725
	NG v Milk		1
	UK4 v Milk		0.801
Colon 5HIAA:5HT ratio	Fr1 v Milk	Ratio	<0.001
	NG v Milk		0.088
	UK4 v Milk		0.94923
Colon 5HT	Fr1 v Milk	5HT (µg/g tissue)	0.158
	NG v Milk		0.066
	UK4 v Milk		0.266
Colon length	Fr1 v Milk	Colong length (cm)	0.893
	NG v Milk		1
	UK4 v Milk		0.917
Elevate plus maze	Fr1 v Milk	Time spent in open arm (s)	0.966
	NG v Milk		0.758
	UK4 v Milk		1
Faecal pellet weight	Fr1 v Milk	Weight per pellet (g)	1
	NG v Milk		0.943
	UK4 v Milk		0.986
Faecal water content	Fr1 v Milk	Faecal water content (%)	1
	NG v Milk		0.892
	UK4 v Milk		0.884
Fat mass	Fr1 v Milk	Fat mass (%)	0.982
	NG v Milk		0.816
	UK4 v Milk		0.99
Fear conditioning Phase 1: Acquisition-Context	Fr1 v Milk	End	0.369
		Interval 1	0.797
		Interval 2	0.992

Test	Comparison	Measure	P-value
		Interval 3	0.968
		Interval 4	0.506
		Interval 5	0.663
		Interval 6	0.374
		Start	0.763
	NG v Milk	End	0.577
		Interval 1	0.949
		Interval 2	0.994
		Interval 3	0.604
		Interval 4	0.919
		Interval 5	0.48
		Interval 6	0.22
		Start	0.982
	UK4 v Milk	End	1
		Interval 1	0.455
		Interval 2	0.448
		Interval 3	0.882
		Interval 4	0.649
		Interval 5	0.89
		Interval 6	1
		Start	0.99
Fear conditioning Phase 1: Acquisition-Cue	Fr1 v Milk	Cue 1	0.52
		Cue 2	0.843
		Cue 3	0.6
		Cue 4	0.6
		Cue 5	0.0991
		Cue 6	0.915
	NG v Milk	Cue 1	0.985
		Cue 2	0.961
		Cue 3	0.941
		Cue 4	0.941
		Cue 5	0.1091
		Cue 6	0.58
	UK4 v Milk	Cue 1	0.899
		Cue 2	0.677
		Cue 3	0.767
		Cue 4	0.767
		Cue 5	0.6183
		Cue 6	1
Fear Conditioning Phase 2 - Contextual learning	NG v Milk	Percentage freezing (%)	0.2363
	UK4 v Milk		0.055
	Fr1 v Milk		0.1095
Fear conditioning Phase 2: Cued learning	Fr1 v Milk	Presentations of the cue: D1	0.999

Test	Comparison	Measure	P-value
		Presentations of the cue: D10	0.999
		Presentations of the cue: D2	0.99
		Presentations of the cue: D3	0.785
		Presentations of the cue: D4	0.601
		Presentations of the cue: D5	0.999
		Presentations of the cue: D6	0.989
		Presentations of the cue: D7	0.289
		Presentations of the cue: D8	0.967
		Presentations of the cue: D9	0.938
	NG v Milk	Presentations of the cue: D1	0.563
		Presentations of the cue: D10	0.996
		Presentations of the cue: D2	0.359
		Presentations of the cue: D3	0.16
		Presentations of the cue: D4	0.201
		Presentations of the cue: D5	0.902
		Presentations of the cue: D6	0.469
		Presentations of the cue: D7	0.562
		Presentations of the cue: D8	0.998
		Presentations of the cue: D9	0.898
	UK4 v Milk	Presentations of the cue: D1	0.327
		Presentations of the cue: D10	0.954
		Presentations of the cue: D2	0.216
		Presentations of the cue: D3	0.344
		Presentations of the cue: D4	0.823
		Presentations of the cue: D5	0.807
		Presentations of the cue: D6	0.344
		Presentations of the cue: D7	0.465
		Presentations of the cue: D8	0.953
		Presentations of the cue: D9	0.992
Female urine sniffing test	Fr1 v Milk	Interaction time (s): Urine	0.1695
		Interaction time (s): Water	0.0189
	NG v Milk	Interaction time (s): Water	0.012
		Interaction time (s): Urine	0.2942
	UK4 v Milk	Interaction time (s): Urine	0.043
		Interaction time (s): Water	0.2522
Fluid mass	Fr1 v Milk	Fluid mass (%)	0.584
	NG v Milk		1
	UK4 v Milk		0.83
Food intake	Fr1 v Milk	Food intake (g)	0.202
	NG v Milk		0.977
	UK4 v Milk		0.481
Forced swim test	Fr1 v Milk	Immobility time (s)	0.4062
	NG v Milk		0.6487

Test	Comparison	Measure	P-value
	UK4 v Milk		0.114
Gastrointestinal moltility	Fr1 v Milk	Transit time (s)	0.629
	NG v Milk		1
	UK4 v Milk		0.827
Ileum 5HIAA	Fr1 v Milk	5HIAA (µg/g tissue)	0.948
	NG v Milk		0.266
	UK4 v Milk		0.696
Ileum 5HIAA:5HT ratio	Fr1 v Milk	Ratio	0.99884
	NG v Milk		<0.001
	UK4 v Milk		0.99389
Ileum 5HT	Fr1 v Milk	5HT (µg/g tissue)	0.9079
	NG v Milk		0.015
	UK4 v Milk		0.8529
Lean mass	Fr1 v Milk	Lean mass (%)	0.99
	NG v Milk		0.821
	UK4 v Milk		0.999
Marble burying test	Fr1 v Milk	Marbles buried	0.5191
	NG v Milk		0.83514
	UK4 v Milk		0.009
Mesenteric lymph nodes pTreg	Fr1 v Milk	Percentage CD4+, CD25+, FoxP3+, Helios-cells	0.7727
	NG v Milk		0.9877
	UK4 v Milk		0.002
Mesenteric lymph nodes Treg	Fr1 v Milk	Percentage CD4+, CD25+, FoxP3+ cells	0.7728
	NG v Milk		0.9877
	UK4 v Milk		0.001
Open field - Locomotor activity	Fr1 v Milk	Distance moved (cm)	0.998
	NG v Milk		0.848
	UK4 v Milk		1
Open field test	Fr1 v Milk	Time spent in centre (s)	0.882
	NG v Milk		0.679
	UK4 v Milk		0.995
Saccharin preference test	Fr1 v Milk	Saccharin preference (%): 36h	0.002
		Saccharin preference (%): 12h	0.35
		Saccharin preference (%): 24h	0.63
		Saccharin preference (%): 48h	0.211
	NG v Milk	Saccharin preference (%): 12h	0.63
		Saccharin preference (%): 24h	0.66
		Saccharin preference (%): 36h	0.7125
		Saccharin preference (%): 48h	0.885
	UK4 v Milk	Saccharin preference (%): 12h	0.27
		Saccharin preference (%): 24h	0.93
		Saccharin preference (%): 36h	0.4552
		Saccharin preference (%): 48h	0.755

Test	Comparison	Measure	P-value
Stress-induced hyperthermia	Fr1 v Milk	Δ Body temperature (°C)	0.992
	NG v Milk		0.291
	UK4 v Milk		1
Tail-suspension test	Fr1 v Milk	Time spent immobile (s)	0.588
	NG v Milk		0.223
	UK4 v Milk		0.998
Water intake	Fr1 v Milk	Water intake (mL)	0.261
	NG v Milk		0.999
	UK4 v Milk		0.845

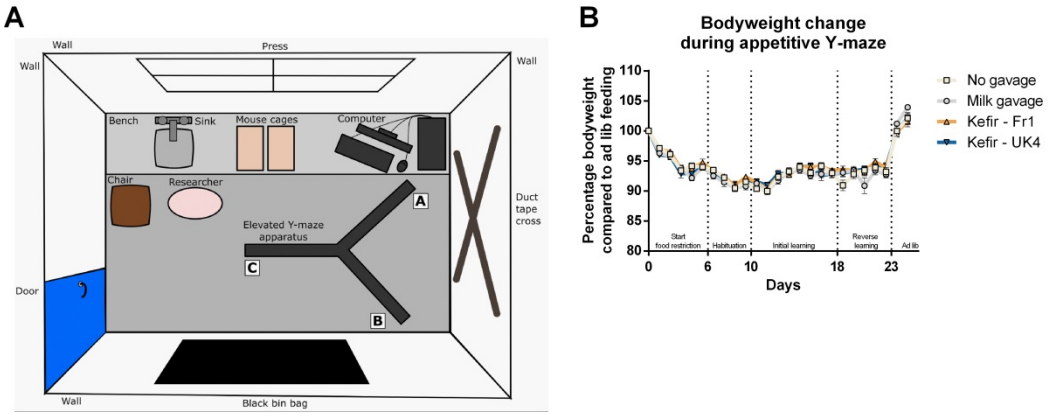


Figure S1: Room layout with cues for the appetitive Y-maze and food restriction. The room layout with the various cues used in the appetitive Y-maze is depicted (A). In addition, mice were kept on food restriction of 90-95% of the free-feeding body weight. All data are expressed as mean ± SEM (n = 12).

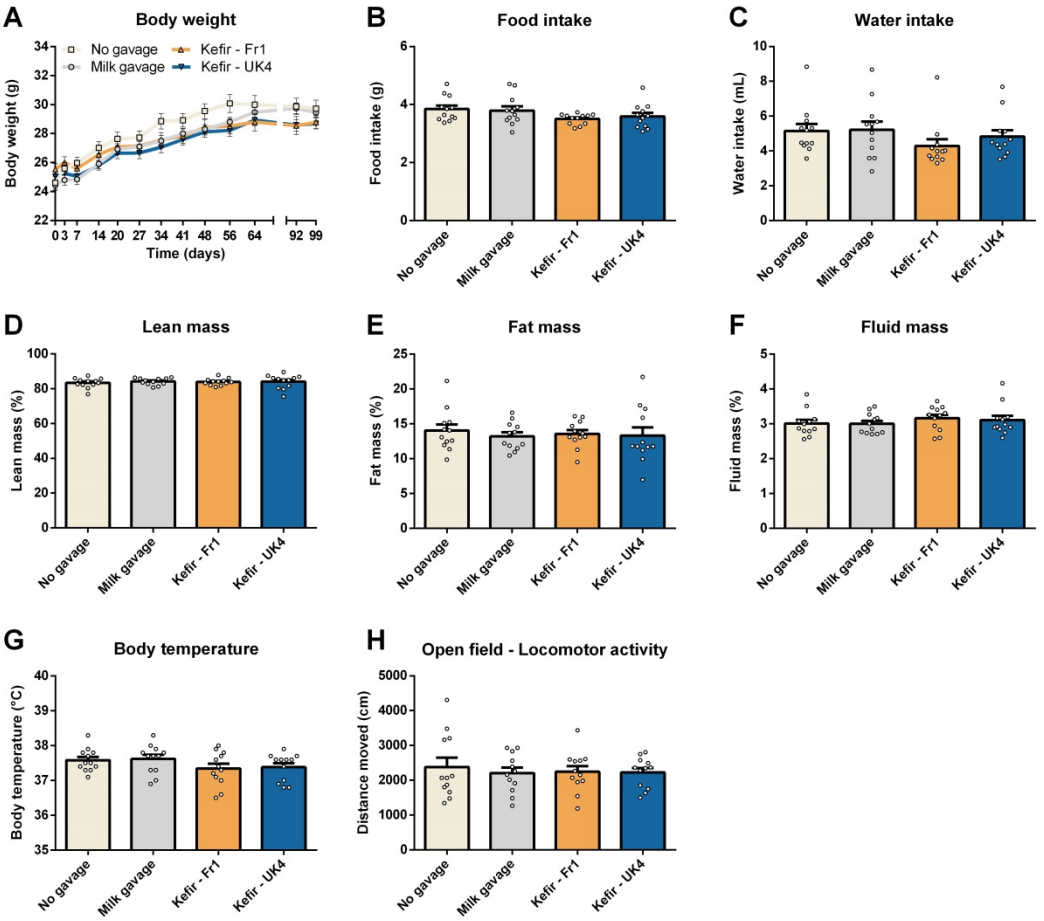


Figure S2: Kefir was well-tolerated. Body weight as measured throughout the study (A). The gap in-between day 64 and 92 represents the appetitive Y-maze, in which animals were food restricted. Food intake and drinking water intake were measured during the habituation phase of the saccharin preference test (B, C). Body composition (i.e. lean, fat and fluid mass) were quantified at the end of the study (D-F). Basal body temperature was taken during the stress-induced hyperthermia test (G). Locomotor activity was assessed in the open field test. All data are expressed as mean ± SEM (n = 11-12). Dots on each graph represent individual animals.

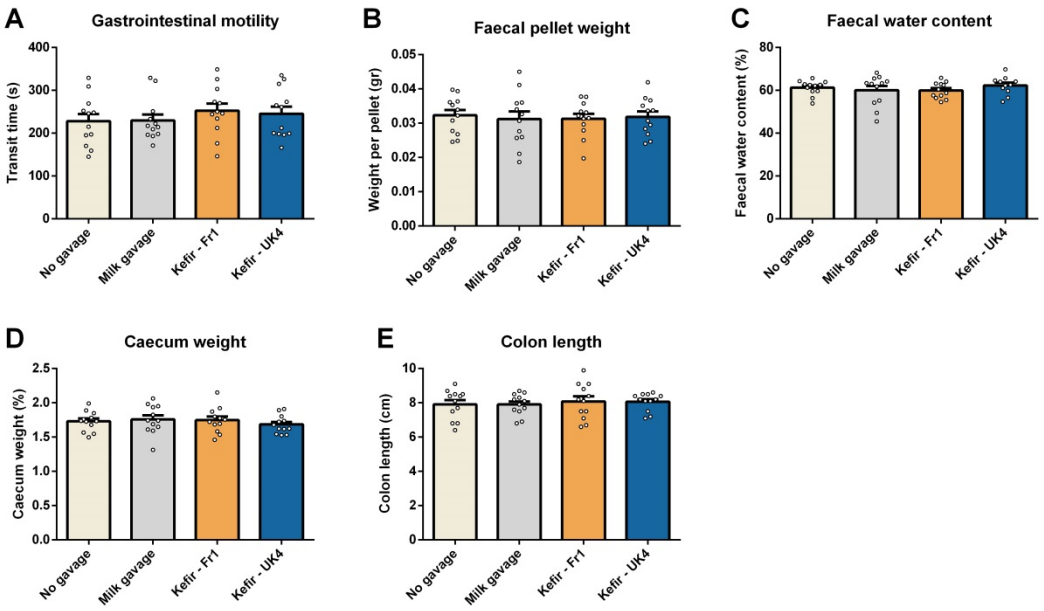


Figure S3: Kefir did not influence gastrointestinal motility. Gastrointestinal motility was assessed by carmine red administration (A). Faecal pellet weight and water content were quantified during the "faecal water content assessment" (B, C). Caecum weight and colon length were measured at the end of the study (D, E). All data are expressed as mean ± SEM (n = 11-12). Dots on each graph represent individual animals.

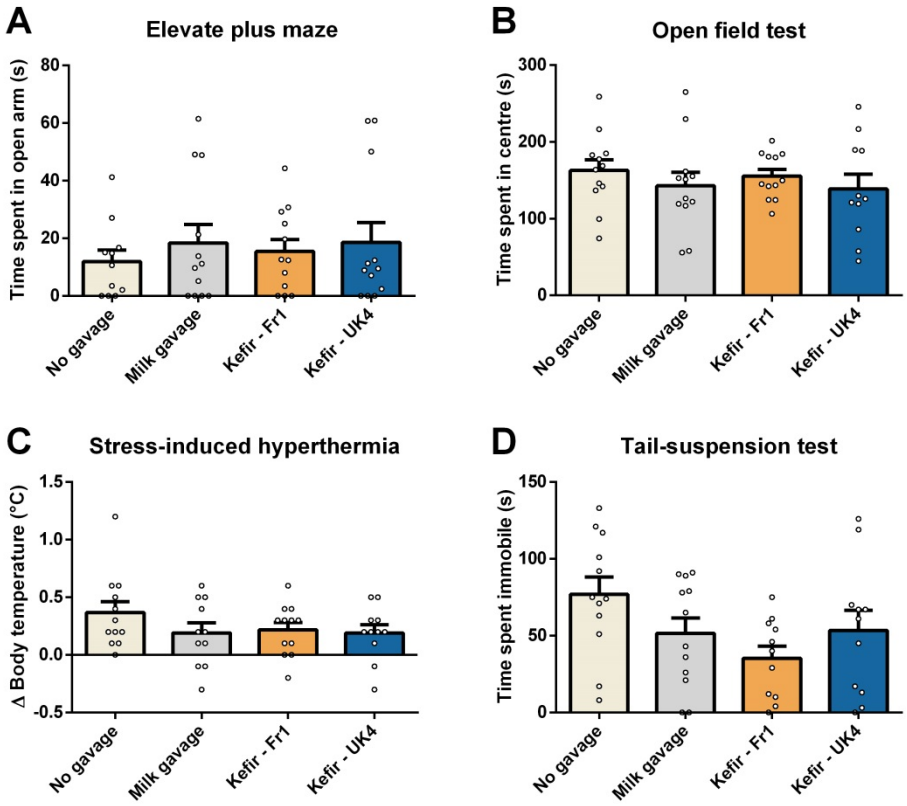


Figure S4: Selective anxiety-like and depressive-like behavioural measurement showed no differences. Repetitive/anxiety-like behaviour was assessed using the elevated plus maze and open field test (A, B). Stress-responsiveness was determined using the stress-induced hyperthermia test (C). Depressive-like behaviour was investigated using the tail suspension test (D). All data are expressed as mean ± SEM (n = 11-12). Dots on each graph represent individual animals.

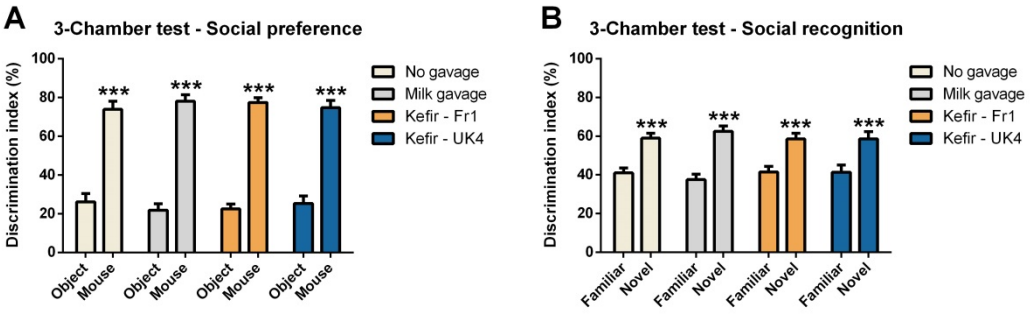


Figure S5: Kefir did not influence social preference or recognition. Social preference and recognition were assessed with the 3-chamber social interaction test (A, B). All data are expressed as mean ± SEM (n = 12).

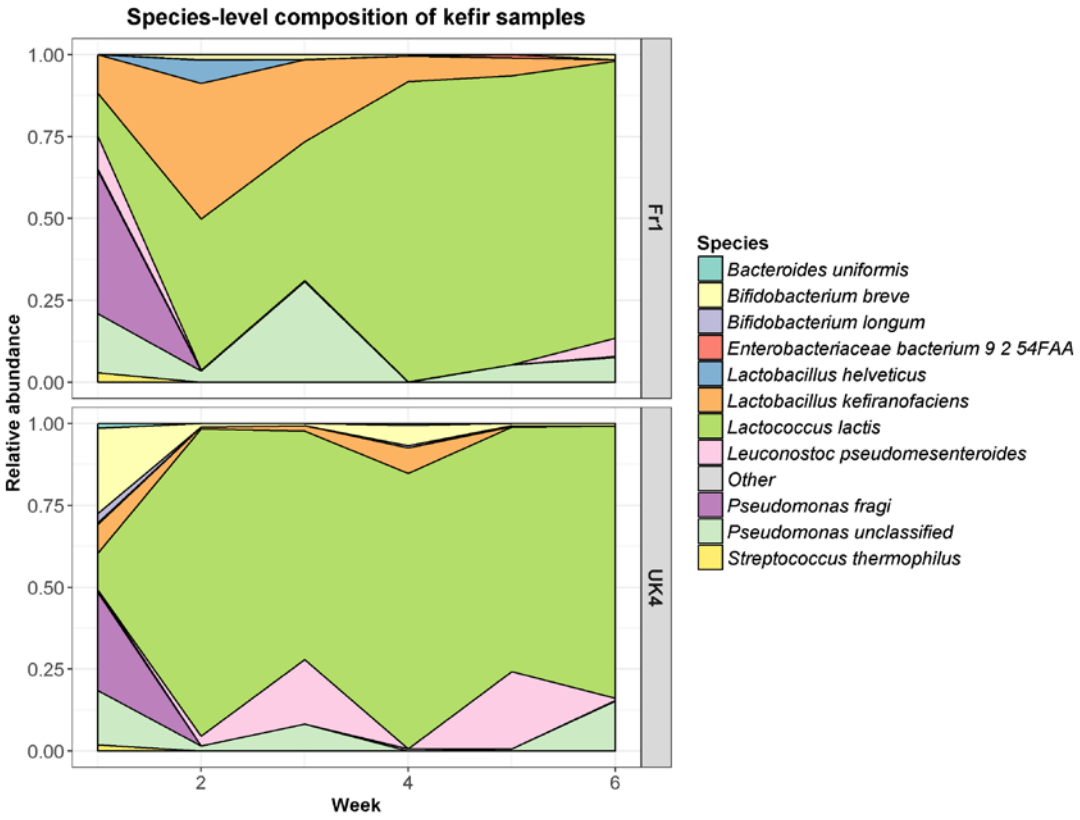


Figure S6: Stacked area chart showing the microbial composition of kefirs over the course of the experiment.

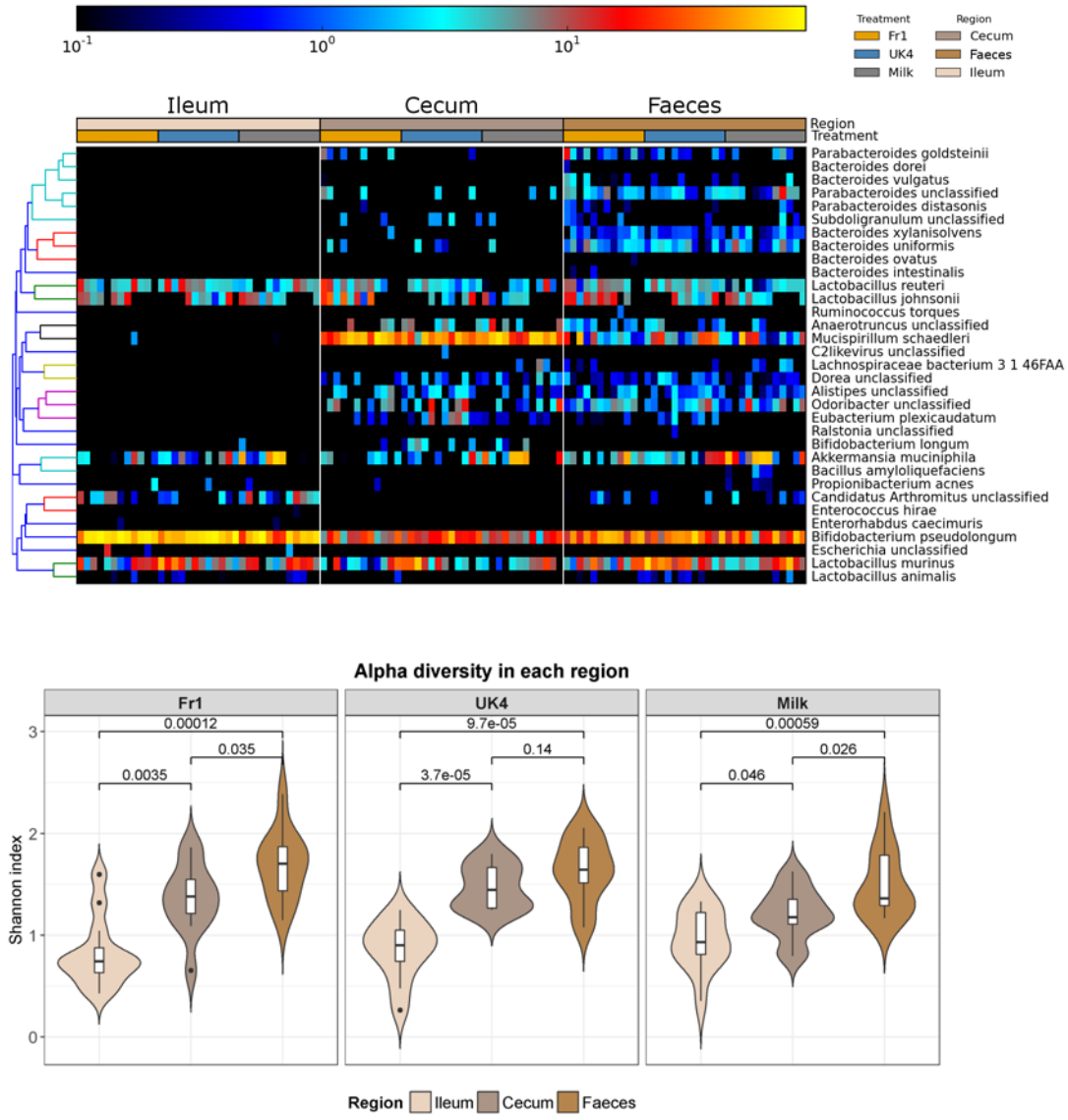


Figure S7: Compositional analysis of the murine gastrointestinal (GI) tract within each group. (A) Heatmap showing the 25 most abundant species across each region of the GI tract. (B) Violin plots showing differences in alpha diversity across each GI region.

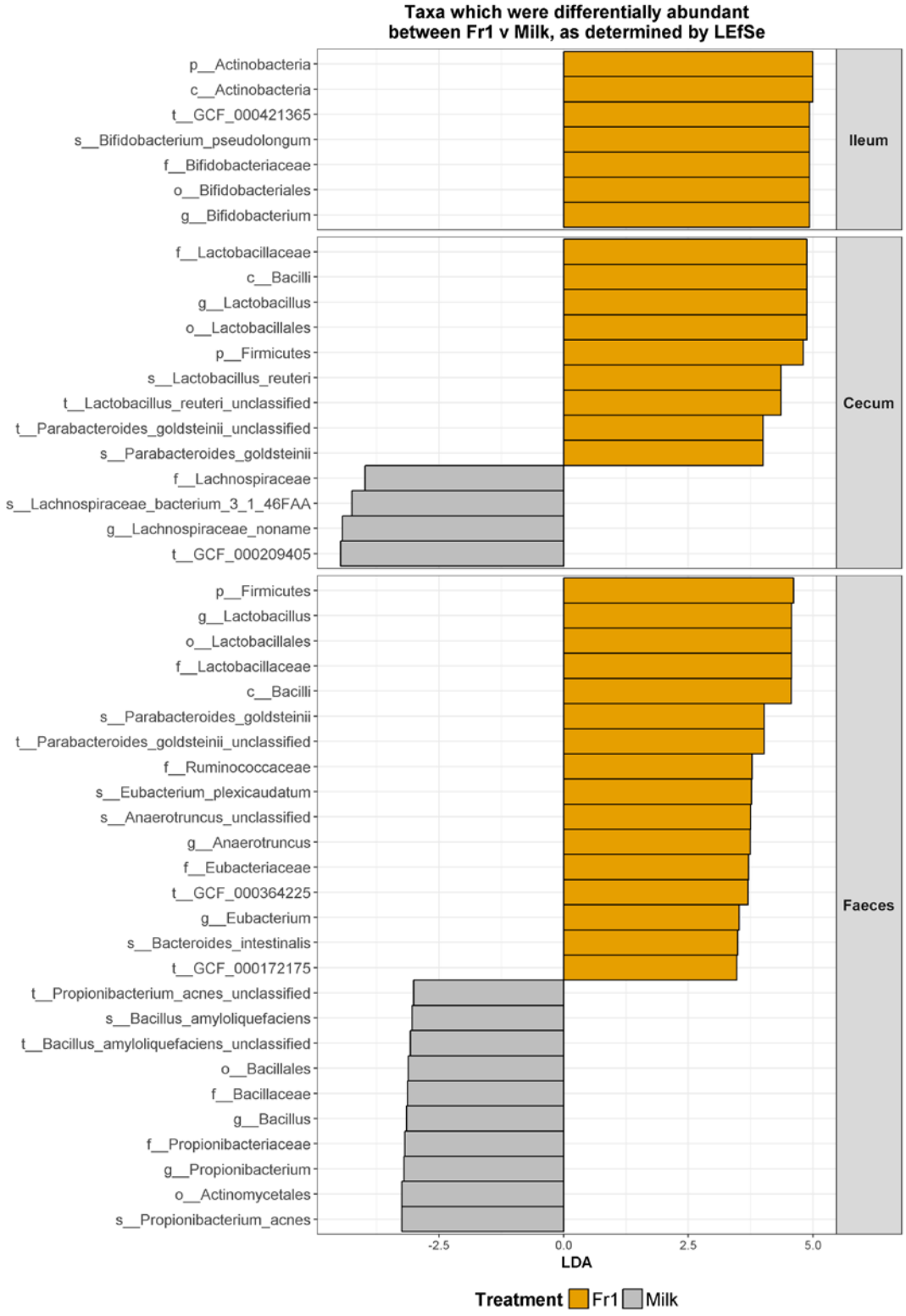


Figure S8: Taxa which were differentially abundant between Fr1 versus Milk-fed mice, as determined by LEfSe.

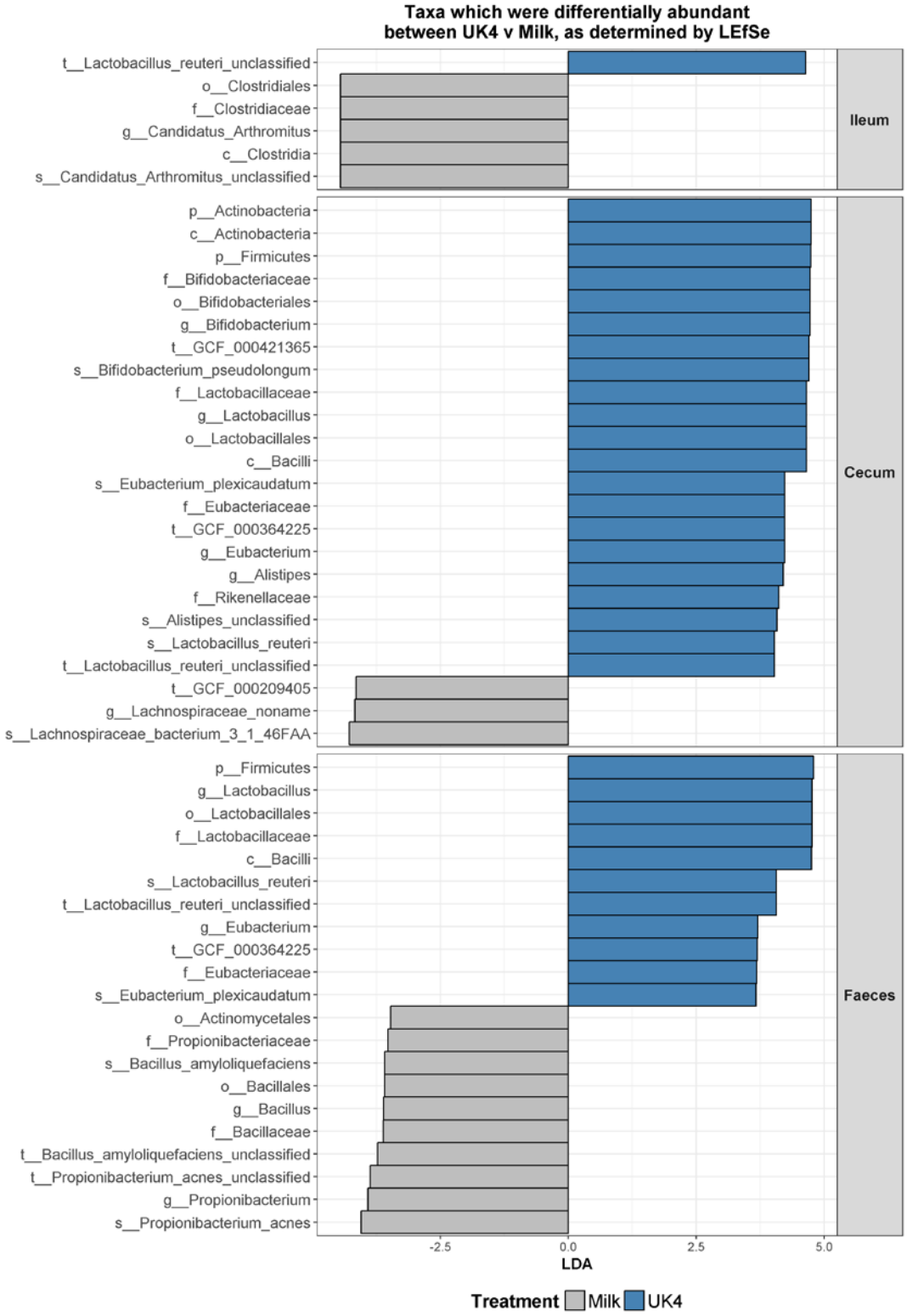


Figure S9: Taxa which were differentially abundant between UK4 versus Milk-fed mice, as determined by LEfSe.

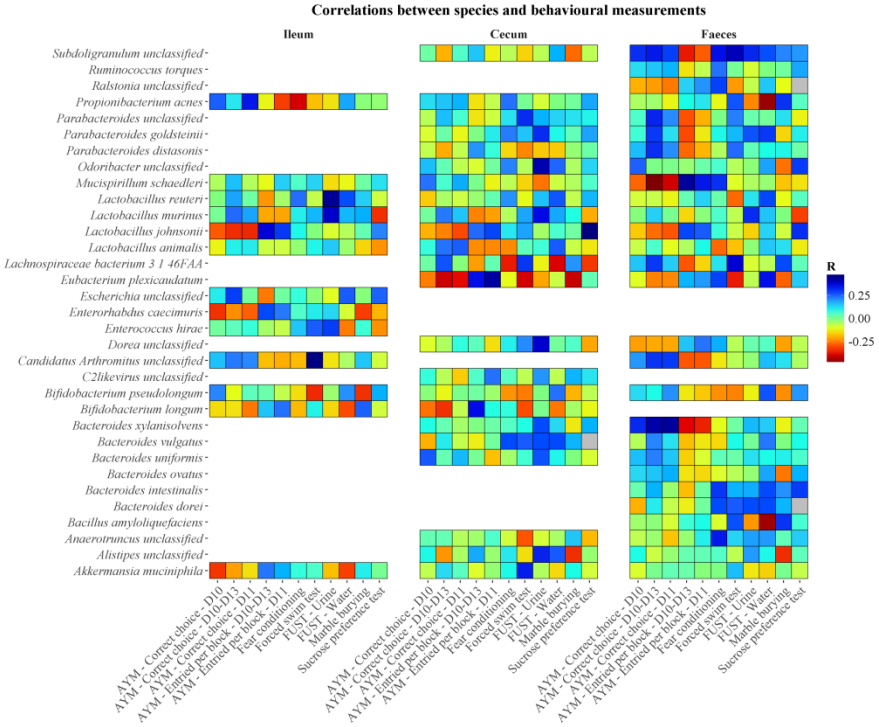


Figure S10: Correlations between species and immuno-physiological parameters. The heatmap shows the Spearman rank correlation coefficient for each combination of variables. HAllA indicated that none of these correlations were significant.

Table S2: Reference genomes which were included in the custom PanPhlAn pangenome databases used in this study.

Species	Assembly	Strain
Lactobacillus reuteri	GCF_000010005	JCM 1112
Lactobacillus reuteri	GCF_000016825	DSM 20016
Lactobacillus reuteri	GCF_000159455	SD2112
Lactobacillus reuteri	GCF_000236455	ATCC 53608
Lactobacillus reuteri	GCF_000410995	I5007
Lactobacillus reuteri	GCF_000439275	TD1
Lactobacillus reuteri	GCF_001046835	IRT
Lactobacillus reuteri	GCF_001618905	ZLR003
Lactobacillus reuteri	GCF_001688685	I49
Bifidobacterium pseudolongum	GCF_000800475	PV8-2
Bifidobacterium pseudolongum	GCF_002282915	UMB-MBP-01
Bifidobacterium pseudolongum	GCF_002706665	DSM 20092

Table S3: Enzyme Commission (EC) level 4 categories which were differentially abundant between kefir versus Milk-fed mice, as determined by LEfSe.

Comparison	Enzyme	Accepted name	Reaction catalysed	Treatment	LDA	Alpha
Cecum: Fr1 v Milk	1.1.1.27	L-lactate dehydrogenase	(S)-lactate + NAD(+) <=> pyruvate + NADH	Fr1	3.219828704	0.01655138
Cecum: Fr1 v Milk	2.4.2.8	Hypoxanthine phosphoribosyltransferase	IMP + diphosphate <=> hypoxanthine + 5-phospho-alpha-D-ribose 1-diphosphate	Fr1	3.141821655	0.021052562
Cecum: Fr1 v Milk	5.4.2.11	Phosphoglycerate mutase (2,3-diphosphoglycerate-dependent)	2-phospho-D-glycerate <=> 3-phospho-D-glycerate	Fr1	3.099165012	0.037500712
Cecum: Fr1 v Milk	1.5.1.36	Flavin reductase (NADH)	Reduced flavin + NAD(+) <=> flavin + NADH	Fr1	3.086778546	0.019050703
Cecum: Fr1 v Milk	2.3.1.n3	Glycerol-3-phosphate acyltransferase (acyl-phosphate transferring)	Acyl-phosphate + sn-glycerol 3-phosphate <=> 1-acyl-sn-glycerol 3-phosphate + phosphate	Fr1	3.002299062	0.036169089
Ileum: Fr1 v Milk	2.6.1.16	Glutamine--fructose-6-phosphate transaminase (isomerizing)	L-glutamine + D-fructose 6-phosphate <=> L-glutamate + D-glucosamine 6-phosphate	Fr1	3.338367894	0.001822735
Ileum: Fr1 v Milk	6.3.1.2	Glutamate--ammonia ligase	ATP + L-glutamate + NH(3) <=> ADP + phosphate + L-glutamine	Fr1	3.471479978	0.024343384
Ileum: Fr1 v Milk	4.3.1.19	Threonine ammonia-lyase	L-threonine <=> 2-oxobutanoate + NH(3)	Fr1	3.313442872	0.017926339
Ileum: Fr1 v Milk	5.1.3.4	L-ribulose-5-phosphate 4-epimerase	L-ribulose 5-phosphate <=> D-xylulose 5-phosphate	Fr1	3.597203657	0.01304252
Ileum: Fr1 v Milk	3.2.1.22	Alpha-galactosidase	Hydrolysis of terminal, non-reducing alpha-D-galactose residues in alpha-D-galactosides, including galactose oligosaccharides, galactomannans and galactolipids	Fr1	3.502049323	0.043308143
Ileum: Fr1 v Milk	2.4.2.9	Uracil phosphoribosyltransferase	UMP + diphosphate <=> uracil + 5-phospho-alpha-D-ribose 1-diphosphate	Fr1	3.496299022	0.011074438
Ileum: Fr1 v Milk	3.2.1.21	Beta-glucosidase	Hydrolysis of terminal, non-reducing beta-D-glucosyl residues with release of beta-D-glucose	Fr1	3.377517348	0.024343384
Ileum: Fr1 v Milk	3.2.1.38	Beta-D-fucosidase	Hydrolysis of terminal non-reducing beta-D-fucose residues in beta-D-fucosides	Fr1	3.377517348	0.024343384
Ileum: Fr1 v Milk	1.1.1.274	2,5-didehydrogluconate reductase (2-dehydro-D-gluconate-forming)	2-dehydro-D-gluconate + NADP(+) <=> 2,5-didehydro-D-gluconate + NADPH	Fr1	3.376891932	0.043308143
Ileum: Fr1 v Milk	5.3.1.4	L-arabinose isomerase	L-arabinose <=> L-ribulose	Fr1	3.352503195	0.009374768
Ileum: Fr1 v Milk	5.4.2.2	Phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent)	Alpha-D-glucose 1-phosphate <=> alpha-D-glucose 6-phosphate	Fr1	3.316771064	0.043308143
Ileum: Fr1 v Milk	2.6.1.42	Branched-chain-amino-acid transaminase	L-leucine + 2-oxoglutarate <=> 4-methyl-2-oxopentanoate + L-glutamate	Fr1	3.289575919	0.037666922
Ileum: Fr1 v Milk	1.1.1.27	L-lactate dehydrogenase	(S)-lactate + NAD(+) <=> pyruvate + NADH	Fr1	3.284127788	0.006656727
Ileum: Fr1 v Milk	4.2.1.11	Phosphopyruvate hydratase	2-phospho-D-glycerate <=> phosphoenolpyruvate + H(2)O	Fr1	3.209400442	0.011074438
Ileum: Fr1 v Milk	1.2.1.11	Aspartate-semialdehyde dehydrogenase	L-aspartate 4-semialdehyde + phosphate + NADP(+) <=> L-4-aspartyl phosphate + NADPH	Fr1	3.13617197	0.037666922
Ileum: Fr1 v Milk	5.2.1.8	Peptidylprolyl isomerase	Peptidylproline (omega=180) <=> peptidylproline (omega=0)	Milk	3.311774084	0.04964723
Ileum: Fr1 v Milk	4.2.99.18	DNA-(apurinic or apyrimidinic site) lyase	The C-O-P bond 3' to the apurinic or apyrimidinic site in DNA is broken by a beta-elimination reaction, leaving a 3'-terminal unsaturated sugar and a product with a terminal 5'-phosphate	Milk	3.029327569	0.028103306
Stool: Fr1 v Milk	3.6.3.27	Phosphate-transporting ATPase	ATP + H(2)O + phosphate(Out) <=> ADP + phosphate + phosphate(In)	Fr1	3.391469552	0.009374768
Cecum: UK4 v Milk	4.2.1.20	Tryptophan synthase	L-serine + 1-C-(indol-3-yl)glycerol 3-phosphate <=> L-tryptophan + D-glyceraldehyde 3-phosphate + H(2)O	Milk	3.028761616	0.028240369
Cecum: UK4 v Milk	4.2.99.18	DNA-(apurinic or apyrimidinic site) lyase	The C-O-P bond 3' to the apurinic or apyrimidinic site in DNA is broken by a beta-elimination reaction, leaving a 3'-terminal unsaturated sugar and a product with a terminal 5'-phosphate	Milk	3.028359552	0.022546702
Cecum: UK4 v Milk	6.3.2.2	Glutamate--cysteine ligase	ATP + L-glutamate + L-cysteine <=> ADP + phosphate + gamma-L-glutamyl-L-cysteine	UK4	3.147997304	0.037666922
Cecum: UK4 v Milk	3.6.3.27	Phosphate-transporting ATPase	ATP + H(2)O + phosphate(Out) <=> ADP + phosphate + phosphate(In)	UK4	3.744720001	0.032625612
Cecum: UK4 v Milk	2.4.2.9	Uracil phosphoribosyltransferase	UMP + diphosphate <=> uracil + 5-phospho-alpha-D-ribose 1-diphosphate	UK4	3.120486959	0.003228225
Cecum: UK4 v Milk	6.3.2.6	Phosphoribosylaminoimidazolesuccinocarboxamide synthase	ATP + 5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate + L-aspartate <=> ADP + phosphate + (S)-2-(5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamido)succinate	UK4	3.014603844	0.043308143
Ileum: UK4 v Milk	2.6.1.16	Glutamine--fructose-6-phosphate transaminase (isomerizing)	L-glutamine + D-fructose 6-phosphate <=> L-glutamate + D-glucosamine 6-phosphate	UK4	3.10300258	0.020921335
Ileum: UK4 v Milk	1.2.1.11	Aspartate-semialdehyde dehydrogenase	L-aspartate 4-semialdehyde + phosphate + NADP(+) <=> L-4-aspartyl phosphate + NADPH	UK4	3.179455722	0.043308143
Stool: UK4 v Milk	1.6.5.11	NADH dehydrogenase (quinone)	NADH + a quinone <=> NAD(+) + a quinol	Milk	3.390974785	0.032663442
Stool: UK4 v Milk	6.3.2.2	Glutamate--cysteine ligase	ATP + L-glutamate + L-cysteine <=> ADP + phosphate + gamma-L-glutamyl-L-cysteine	UK4	3.028164813	0.011074438
Stool: UK4 v Milk	3.6.3.27	Phosphate-transporting ATPase	ATP + H(2)O + phosphate(Out) <=> ADP + phosphate + phosphate(In)	UK4	3.458753884	0.003892417
Stool: UK4 v Milk	2.7.7.6	DNA-directed RNA polymerase	Nucleoside triphosphate + RNA(n) <=> diphosphate + RNA(n+1)	UK4	3.284084737	0.020921335
Stool: UK4 v Milk	3.5.1.5	Urease	Urea + H(2)O <=> CO(2) + 2 NH(3)	UK4	3.167115479	0.007911789
Stool: UK4 v Milk	2.7.2.1	Acetate kinase	ATP + acetate <=> ADP + acetyl phosphate	UK4	3.094500366	0.01304252
Stool: UK4 v Milk	2.7.7.7	DNA-directed DNA polymerase	Deoxynucleoside triphosphate + DNA(n) <=> diphosphate + DNA(n+1)	UK4	3.025162116	0.024343384

General Discussion

As discussed in Chapter 1, the field of food microbiology has been revolutionised by the advent of high-throughput sequencing (HTS). This technology enables unprecedented characterisation of food-related microbial isolates, including starter cultures, probiotics, and foodborne pathogens. Additionally, and of particular relevance to this thesis, HTS allows culture-independent metagenomic analysis of the mixed microbial communities, or microbiota, present in fermented foods. As stated in Chapter 2, food fermentation has been practised for millennia as a means to preserve or enhance foods. Today, fermented foods are becoming increasingly popular since many health benefits, including anti-diabetic, anti-inflammatory, and anti-obesity effects, have been attributed to them (1). HTS has been extensively utilised to catalogue the microbial compositions of an array of fermented foods, but it can also be employed to predict or measure the activities of microbes during fermentations, which yields insights into microbial dynamics *in situ.* Such information may shed light on the ways in which microbes contribute to qualities such as flavour in fermented foods, and thus it might be used to optimise fermentations to produce food with desired properties. Another important consideration for producing fermented foods is safety, and, as mentioned in Chapter 1, HTS may potentially be applied to detect pathogens in these foods. Furthermore, in Chapter 1, we also highlight that HTS might be used to determine the effects of fermented foods on the gut microbiota, which may help to elucidate the underlying mechanisms responsible for the health benefits associated with these foods. In this thesis, we have demonstrated that HTS, especially shotgun metagenomics, is an invaluable tool to (i) expand our understanding on the microbiology of food fermentations, (ii) ensure the safety of fermented foods, and (iii) investigate their impact on the host.

In Chapters 3 and 4, we investigated the ways in which microbes may influence flavour development in fermented foods. Firstly, in Chapter 3, we utilised shotgun metagenomics to characterise the kefir microbiome during fermentation. Specifically, we examined kefirs from three separate countries. We observed consistent patterns in microbial succession in the analysed kefirs, and, additionally, we found that changes in the microbiota corresponded with changes in the metabolome. Notably, we observed that particular species correlated with particular flavour compounds, which suggested that the different microbes present in kefir had distinct effects on its flavour. Indeed, we subsequently confirmed that spiking milk with isolates from kefir resulted in predictable changes in flavour compounds. A similar approach was taken in Chapter 4 to characterise smear ripened cheeses during ripening, where we again observed that particular microbes correlated with particular flavour compounds. Importantly, we detected pathways associated with flavour development in both studies. Our work highlights that the microbiota is, unsurprisingly, linked to flavour development in fermented food. Crucially, it also reveals that sequencing can be used to understand the ways in which microbes contribute to flavour in fermented foods. We propose that such knowledge might ultimately be used to design starter mixes to produce fermented foods with enhanced flavours. Future work will focus on characterising microbial gene expression during food fermentations to gain deeper insights into the intricate networks through which microbes contribute to flavour (2, 3). Alternatively, metagenome-scale metabolic modelling (4) is another approach which may enable us to predict *in silico* the flavour compounds produced by starters.

Safety, rather than flavour, is undoubtedly the most important food quality, and foodborne pathogens are responsible for millions of illnesses, annually (5). In

Chapter 5, we assess the potential to use shotgun metagenomics to detect pathogens in fermented foods. A previous study had already achieved strain-level detection of pathogens in spinach samples that had been spiked with *Escherichia coli*, but the methods used therein relied on metagenome assembly (6), which is a computationally intensive process not conducive to rapid large-scale testing, and thus alternative approaches are desirable. In Chapter 5, we addressed this issue by demonstrating that three short-read alignment-based tools, MetaMLST (7), PanPhlAn (8), and StrainPhlAn (9), accurately and rapidly, detected pathogens in the aforementioned spinach samples. Subsequently, we employed these tools to test the safety of nunu, which is a traditional fermented dairy beverage from Ghana that is produced by the spontaneous fermentation of raw cow milk. We observed that nunu was frequently contaminated with gut-associated bacteria, and, worryingly, we detected putatively pathogenic strains in several samples. Thus, we concluded that better hygiene practises were imperative for producing safer nunu. Overall, these results show that short-read alignment approaches may be suitable food safety tools, and we expect that they can be applied to inform safety measures during production, detect pathogens in products, or trace outbreaks to source. We envisage that this technology may eventually be adopted by the food industry, especially if sequencing costs continue to decrease. The ability to simultaneously detect every pathogen in a fermented food is, unquestionably, invaluable.

In Chapters 3 to 5 we highlighted several examples of the ways in which shotgun metagenomics can be applied to characterise fermented food microbiota. However, currently, there is still no consensus on the optimal methods to use for such analyses. Therefore, in Chapter 6, we investigated the influences of sequencer choice, sequencing depth and bioinformatics methodologies on the analysis of fermented

food metagenomes. We found that three high-throughput short-read sequencers, the Illumina MiSeq, NextSeq 500, and Ion Proton, provided accordant results at divergent sequencing depths. Functional analysis with SUPER-FOCUS (10) gave congruent results across the sequencers at sequencing depths ranging from 100,000 to 7,500,000 reads. Strain-level analysis with PanPhlAn produced similar results across the sequencers, and, remarkably, it correctly identified the dominant strains in every kefir sample with over 500,000 reads. Compositional analysis, with a given species classifier, was also accordant across the sequencers at different sequencing depths. However, the compositional results from classifiers were significantly different to each other. Notably, we observed that the species abundances predicted by each classifier apart from MetaPhlAn2 (11) were biased by their respective reference genome sizes. Furthermore, we identified different false positive rates between the classifiers, of which MetaPhlAn2 produced the fewest false positive results. Thus, our findings highlight that species classifier choice is pivotally important when analysing fermented food metagenomes, and they suggest that MetaPhlAn2 is perhaps the most accurate species classifier analysed here. Additionally, there is a preconception that shotgun metagenomics requires a considerable sequencing depth per sample (12), but our work suggests that this is not necessarily true for fermented food microbiota. Indeed, we found that over 500,000 reads per sample was sufficient even for strain-level analysis. We hope that this study will guide other food microbiologists in their efforts to design shotgun metagenomics experiments.

While in Chapters 3-6 we used shotgun metagenomics to characterise fermented food microbiota; in Chapter 7 we examined if consumption of kefir could impact the gut microbiota of mice as this may be a mechanism by which kefir exerts its reported

health benefits (13, 14). To this end, we utilised shotgun metagenomics as a tool; specifically, we combined shotgun metagenomics with behavioural analysis to determine if kefirs, compared to unfermented milk, altered the gut-brain axis in mice. The gut-brain axis refers to bidirectional communication from the gut to the brain, and growing evidence indicates that the gastrointestinal microbiota can influence mood via this system (15, 16). Our work was motivated by recent evidence which suggested that fermented foods can alleviate anxiety or depression (17, 18). Excitingly, we observed that kefirs modulated the gut microbiota in mice, while they simultaneously ameliorated anxious and depressive-like behaviours in the animals. We used species-level analysis alongside strain-level functional analysis to reveal that kefir ingestion induced an apparent increase in the relative abundance of bacteria containing genes for gamma-aminobutyric acid (GABA) production along with tryptophan biosynthesis. GABA is the main inhibitory neurotransmitter in the central nervous system (19), while tryptophan is a precursor to serotonin. Importantly, deficiencies in both have been linked to anxiety and depression (20, 21). Thus, our discoveries hint that kefir reduced anxious and depressive-like behaviours by increasing the capacity for the gut microbiome to synthesise these neurotransmitters. However, more work, such as host gene expression analysis, metabolomics, or metatranscriptomics, is needed to discern the mechanism by which kefir exerted these effects in mice. Additionally, trials are necessary to assess if kefir may be used as a dietary intervention for anxiety in humans.

In conclusion, in this thesis, I have demonstrated the value of high-throughput sequencing-based characterisation of fermented foods and their impact on the host gut microbiota.

References

1. **Marco ML, Heeney D, Binda S, Cifelli CJ, Cotter PD, Foligne B, Gänzle M, Kort R, Pasin G, Pihlanto A.** 2017. Health benefits of fermented foods: microbiota and beyond. Current Opinion in Biotechnology **44:**94-102.
2. **De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D.** 2016. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. Scientific Reports **6**.
3. **Song Z, Du H, Zhang Y, Xu Y.** 2017. Unraveling core functional microbiota in traditional solid-state fermentation by high-throughput amplicons and metatranscriptomics sequencing. Frontiers in Microbiology **8:**1294.
4. **Magnúsdóttir S, Thiele I.** 2018. Modeling metabolism of the human gut microbiome. Current Opinion in Biotechnology **51:**90-96.
5. **Scallan E, Hoekstra R, Mahon B, Jones T, Griffin P.** 2015. An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. Epidemiology and Infection **143:**2795-2804.
6. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2016. Strain-Level Discrimination of Shiga Toxin-Producing Escherichia coli in Spinach Using Metagenomic Sequencing. PloS One **11:**e0167870.
7. **Zolfo M, Tett A, Jousson O, Donati C, Segata N.** 2016. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Research**:**gkw837.

8. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods **13:**435-438.

9. **Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N.** 2017. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research **27:**626-638.

10. **Silva GGZ, Green KT, Dutilh BE, Edwards RA.** 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics **32:**354-361.

11. **Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N.** 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods **12:**902-903.

12. **Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R.** 2012. Unlocking the potential of metagenomics through replicated experimental design. Nature Biotechnology **30:**513-520.

13. **Bourrie BCT, Willing BP, Cotter PD.** 2016. The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir. Frontiers in Microbiology **7**.

14. **Rosa DD, Dias MMS, Grześkowiak ŁM, Reis SA, Conceição LL, Peluzio MdCG.** 2017. Milk kefir: nutritional, microbiological and health benefits. Nutrition Research Reviews **30:**82-96.

15. **Cryan JF, Dinan TG.** 2012. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nat Rev Neurosci **13:**701-712.

16. **Mayer EA, Knight R, Mazmanian SK, Cryan JF, Tillisch K.** 2014. Gut Microbes and the Brain: Paradigm Shift in Neuroscience. The Journal of Neuroscience **34:**15490.

17. **Hilimire MR, DeVylder JE, Forestell CA.** 2015. Fermented foods, neuroticism, and social anxiety: An interaction model. Psychiatry Research **228:**203-208.

18. **Miyake Y, Tanaka K, Okubo H, Sasaki S, Arakawa M.** 2014. Intake of dairy products and calcium and prevalence of depressive symptoms during pregnancy in Japan: a cross-sectional study. BJOG: An International Journal of Obstetrics & Gynaecology **122:**336-343.

19. **Owens DF, Kriegstein AR.** 2002. Is there more to gaba than synaptic inhibition? Nature Reviews Neuroscience **3:**715.

20. **Möhler H.** 2012. The GABA system in anxiety and depression and its therapeutic potential. Neuropharmacology **62:**42-53.

21. **Canli T, Lesch K-P.** 2007. Long story short: the serotonin transporter in emotion regulation and social cognition. Nature Neuroscience **10:**1103.